Optimal Screening of Manipulative Agents via Contests*

Yingkai Li† Xiaoyun Qiu‡

November 7, 2023

Abstract

We study the allocation of multiple homogeneous items to multiple agents with unit demand. The allocation of the items depends only on informative signals that agents can manipulate through costly and wasteful efforts, and monetary transfers are not allowed. Examples of such scenarios include college admission and entities lobbying governments for grants or subsidies. We show that the welfare-maximizing mechanism takes the form of a contest and characterize the optimal contest. We apply our results to settings with a large number of contestants. When the number of agents is large relative to the number of items, although the welfare-maximizing mechanism converges to a winner-take-all contest, the principal's payoff does not. When the numbers of items and agents are both large and proportional to each other, agents with intermediate types receive a random allocation, which reduces the equilibrium level of effort of these agents and of agents with higher types.

Keywords— contests, mechanism design without money, signalling, manipulation, coordination

*We especially thank Asher Wolinsky, Bruno Strulovici, Wojciech Olszewski for advice, helpful conversations and comments. We also thank Dirk Bergemann, Ian Ball, Eddie Dekel, Jeff Ely, Kira Goldner, Carl-Christian Groh, Yingni Guo, Andrei Iakovlev, Annie Liang, Bart Lipman, Brendan Lucier, Deniz Kattwinkel, Joshua Mollner, Kyohei Okumura, Alessandro Pavan, Marcin Pęski, Abhishek Sarkar, James Schummer, Philipp Strack, Matthew Thomas and participants at 34th Stony Brook Game Theory Conference for helpful suggestions, comments and discussions. Yingkai Li thanks Sloan Research Fellowship FG-2019-12378 for financial support.

†Cowles Foundation for Research in Economics, Yale University. Email: yingkai.li@yale.edu

‡Department of Economics, Northwestern University. Email: xiaoyun.qiu@u.northwestern.edu

1 Introduction

In many countries college admission aims at selecting talented students by using test scores (e.g., SAT scores). Several government subsidy programs aim at selecting people with the highest need by using social credit score (e.g., FICO scores). Scientific funding agencies aim at selecting ideas or projects of high quality by using peer review scores. These scenarios can be modeled as contests that select winners based on the ranking with respect to a score which itself is just an unproductive signal but reflects the underlying unobserved attributes of interest. While these scores contain information on agents' abilities, they can be manipulated by agent's costly effort, but not as easily as in cheap talk communication. The obvious problem is that this generates inefficient investment in the wasteful signals. For example, students can fake disability status to gain extra time in tests, which could increase their test scores but does not increase their intellectual abilities (Sansone and Sansone, 2011).[1] Companies can fake their workforce sizes to meet the criteria for legal and financial preferential treatments (Askenazy et al., 2022). In scientific funding, applicants are incentivized to spend effort to overstate the merits of their proposals rather than to develop proposals with high merit (Conix et al., 2021).

The goal of this paper is to characterize the mechanisms that maximize any weighted average between *matching efficiency* and the *sum of agents' utilities*. Our paper provides a reduced form approach to model a screening problem with multiple agents where hidden effort is unproductive. Different from the signaling literature and the recent gaming literature (e.g., Spence, 1973; Frankel and Kartik, 2019; Ball, 2019), the mechanism designer has limited resources in our setting.[2] Implementing full matching efficiency usually encourages undesirable effort that blurs the accuracy of the signals. This competition effect among agents, absent in the signaling and gaming literature, is non-negligible in our setting. In contrast with the contests literature that study competition among agents, the main departure in our setting is that effort is unproductive and purely a social waste. It is well-known from Lazear and Rosen (1981) that organizing a contest better incentivizes effort than piece-rate wage. However, it is unclear whether organizing a contest remains to be a desirable choice in our screening problem where agents can conduct *rent-seeking* act.

Our first result provides a theoretical foundation for the pervasive use of the contest format in reality. In contests, principal chooses a contest rule that is a mapping from signal profile to allocation.[3] This specifies a game and each agent decides an effort strategy in equilibrium. In any

[1] In the school test setting, although exams are usually designed to test students' logical reasoning ability, high scores could sometimes be achieved by "learning stuff by heart" or memorizing past year exams, etc.., which may not contribute to the improvement of logical reasoning ability.

[2] Moreover, unlike the classic signalling setting where it is marginally less costly for high type to exert effort, in our setting, there is no marginal difference on cost across types for one more unit of effort. See Section 3 for more detailed discussion.

[3] Careful readers might wonder what is the proper definition of contest. In Cambridge dictionary, the noun contest is defined as a competition to do better than other people, especially to win a prize or achieve

direct mechanism that implements the equilibrium of a contest, the effort choice and hence the signal recommendation for each agent *only* depends on his own private type, while in a (direct) general mechanism, the signal recommendation to each agent can depend on the whole reported type profile. Therefore, contests rule out the possibility of *coordination* among agents, while in general mechanisms, the principal could elicit information from all agents upfront, and then make signal recommendation to each agent that carefully coordinates their effort. However, we show that these more complex mechanisms are unnecessary: contests are optimal. In particular, we can show that "second-price" format mechanism is strictly dominated by winner-takes-all contest.

The intuition relies on how these mechanisms (dis)incentivize effort, fixing the level of matching efficiency to achieve. In a mechanism that is not a contest, the principal could carefully coordinate agents' effort choices via signal recommendations, which could save effort costs from agents who are not getting any item. Such coordination attempts could reveal information about other agents' types in the signal recommendations, which makes each agent's signal recommendation a lottery that depends on others' types.[4] This makes the following double deviation strategy feasible: agent first misreports his type and later chooses to either (1) follow the signal recommendation when it is favorable to the agent or (2) opt out when it is not favorable. The feasibility of such double deviation strategies increases agents' off-path utilities in mechanisms that are not contests. We further show that this information leakage effect outweighs the efforts saved from coordination when effort cost is not too convex.[5] A contest disables this type of deviation strategies by shutting down the information leakage force coming from the coordination attempt. Hence, fixing the same level of matching efficiency to achieve, agents' off-path deviation utilities for exerting high efforts are weakly[6] lower in contests, resulting in weakly lower effort costs in contests.

Having established that contests are optimal among all mechanisms, we then characterize the optimal contest. Viewed as a direct mechanism, the optimal contest may be described as follows: the type space of each agent is partitioned into three types of intervals under the optimal contest:

a position of leadership or power. However, there is no single unified definition of contests in the economics literature. Contests could have rich prize structures: apart from the simple and commonly used winner-takes-all contests, Moldovanu and Sela (2001) and Olszewski and Siegel (2020) consider contests with multiple and nonidentical prizes. Moreover, contests could be stochastic, e.g. Skaperdas (1996). In computer science literature, contests are modelled as all-pay auctions, which coincide with our definition (see Chawla et al. (2019) and reference therein). In Section 6, we provide a result to show the connection between our contests and the *rank-order contests*. Our contests could be viewed as *coarse ranking* contests.

[4]In comparison, Ben-Porath et al. (2023) consider a resource allocation problem where agents can exert costly effort to acquire evidence that credibly reveals their types. Effort is therefore desirable for the principal in their setting. The optimal mechanism is sequential and features coordination in their model, because it incentivizes most effort.

[5]In the main body of the paper, we assume linear cost function for closed form characterization. In Appendix B, we show that this result holds when cost function is not too convex.

[6]In the formal proof, one can see that any general mechanism that is not a contest is not optimal. In other words, the optimal mechanism has to be a contest.

(1) *no-tension region*: optimal allocation rule coincides with efficient allocation rule and no type in this region exerts effort; (2) *no-effort region*: optimal allocation rule differs from efficient allocation rule and no type in this region has strict incentive to exert effort, and we select the principal-preferred equilibrium where types in no-effort region do not exert effort; and (3) *efficient region*: optimal allocation rule coincides with efficient allocation rule and all types in this region exert positive effort. Intuitively, the principal always wishes to allocate the item efficiently if this does not incentivize effort. However, if implementing the efficient allocation is too lucrative to the agents, this would incur effort. In such scenarios, depending on the weight the principal put on the agents' utilities, either it remains optimal to implement to efficient allocation while inducing effort on-path (*efficient region*), or it is optimal for the principal to selectively choose a fraction of the type space to "flatten" the allocation rule so as to completely eliminate agents' incentives to exert effort (*no-effort region*). We establish that in the optimal contests can only have three categories of regions. However, in general, there could be countably many such intervals, and their order depends on the shape of the type distribution and other primitives.

We apply our results to study contests with a large pool of contestants which we call large contests. We consider two cases. Our first case is when the number of agents is large relative to the number of items. In this case, the optimal contest converges[7] to WTA contest while principal's payoff does not. This setting speaks to applications such as scholarship and research funding, where the number of winners is small relative to the number of applicants. As the number of agents grows large, the efficient allocation rule converges to an allocation rule that is constantly zero except at the highest type. If the item were allocated efficiently, only sufficiently high types would have a non-trivial probability of winning the award and they would have incentives to exert effort. Under the optimal contest, such an incentive is eliminated by choosing a cutoff: if the highest type is below the cutoff, the item is allocated efficiently, while if the highest type is above the cutoff, the item is allocated randomly to all types above the cutoff such that all types above the cutoff have no strict incentive to exert effort. As the number of agents grows, the interval below the cutoff converges to the whole type space. Essentially, the format of contest converges to the winner-takes-all (WTA).

However, principal's payoff under the optimal contest does not converge to that under the winner-takes-all contest, implying that no matter how large the number of contestants is, the principal would not use a WTA contest to approximate the optimal one. Although the measure of the interval above the cutoff is of order $\frac{1}{n}$, its contribution to the principal's objective value is large. This is because the probability that there exists an agent with a type above the cutoff is roughly of order $\frac{1}{e}$. Such an agent would have exerted high effort if it were under the efficient allocation rule, which would have created a large decrease in the agents' utilities and hence principal's objective value. Thus, as long as the principal's payoff has a non-zero weight on agents' utilities, randomly al-

[7]Here convergence means as the number of agents grows, in the optimal contest, the measure of the *no tension* region converges to that of the whole type space.

locating the scarce resource to any agent with sufficiently high signals can substantially improve the objective value. Conceptually, this contrasts with Bulow and Klemperer (1996) who find that the revenue from the efficient allocation rule with one additional buyer exceeds that under the optimal mechanism.

Our second case of a large contest is when the the number of items grows proportionally with the number of agents. This model is better suited for applications such as college admissions, government's financial programs, where a non-negligible fraction of the agents receive some items. In this case, if the items were allocated efficiently, all types above a cutoff would get one item. In the optimal contest, the principal randomizes the allocation for types around this cutoff to eliminate their incentives to exert costly effort, which improves the expected utilities of *middle* types, i.e., types around the cutoff, though at a small cost of lowering matching efficiency. This is reminiscent of the Director's law, which suggests that public programs should be designed primarily to benefit the middle classes. Our finding is also consistent with the empirical finding in Krishna et al. (2022). They use data in Turkey to show that randomly allocating college seats to low score students reduces the stress to all students.

1.1 Related Work

Our paper complements the seminal work of Lazear and Rosen (1981) by providing a different rationale for the adoption of contests in practice. Lazear and Rosen (1981) shows that contests perform better in encouraging effort than piece-rate wage, while our paper shows that the designer cannot do better by considering more general mechanisms. While Lazear and Rosen (1981) advocates that competition is good, our paper shows that only certain amount of competition, rather than the maximal level of competition, is ideal.

Our paper is related to the literature on contests. There is a large literature on characterizing equilibria in contests with various allocation and prize structures (e.g., Barut and Kovenock, 1998; Baye et al., 1993; Che and Gale, 1998; Clark and Riis, 1998; Siegel, 2009, 2010). In these settings, effort is productive and hence desirable for the principal. The effort-maximizing contest has been considered in specific settings, such as those in which principal could only choose whether or not and how to split contestants into sub-contests Moldovanu and Sela (2006), and those in which principal can incentivize more effort by enhancing or reducing inequality in the division of prize among winners (Fang et al., 2020). Zhang (2023) considers the effort-maximizing mechanism and shows that this mechanism could be implemented by a contest due to payoff equivalence. Our paper considers unproductive effort and studies a design problem that does not impose any allocation structure and with a completely different objective, in which the designer aims to allocate items efficiently while also minimizing agents' costs of wasteful efforts. Moreover, we provide a tight characterization of the optimal contest when the number of agents is large, while in the large contests literature (e.g.,

Olszewski and Siegel, 2016, 2020), only the limit of the optimal contests is characterized.

The signals in our model can be viewed as messages sent by the agents to the principal and the corresponding effort costs can be viewed as lying costs. Under this interpretation, the model is different from the literature on partial verification where lying costs are assumes to be either 0 or ∞, depending on the true type and the reported message (e.g., Green and Laffont, 1986). Our model allows a more general form of lying costs, and allows the cost of lying to increase with the magnitude of the lie. In contrast to the literature on evidence (e.g., Ben-Porath et al., 2014; Mylovanov and Zapechelnyuk, 2017), in our model, the evidence is not "hard" because agents can fabricate a signal that is different their true type with some cost.

Similar to the literature on money burning (e.g., Hartline and Roughgarden, 2008; Chawla et al., 2019), agents can exert costly and socially wastefully efforts, which is similar to burning utility. In the money burning literature, the money typically enters agents' utilities in a quasi-linear way that does not depend on agents' types. Moreover, the money burnt is observable. In our model, in contrast, the effort is unobserved. The signal recommendation in our model has to satisfy extra incentive compatibility constraint due to the presence of moral hazard problem in our setting, and the signal enters each agent's utility in a type-dependent way.

Our paper also relates to the literature on signalling games Spence (1973). In signalling games, there is a competitive market that pays each agent a wage that is his estimated type, while in our setting, there is only one designer with a limited budget to allocate among several agents.

Our paper is also related to manipulative behaviors in signalling games (e.g. Frankel and Kartik, 2019; Ball, 2019). The main difference is that we consider a screening setting with multiple agents and limited resources and as a result, the competition effect among agents is a non-negligible force. In those settings and our setting, agent's utility does not satisfy the single crossing property. However, Frankel and Kartik (2019) and Ball (2019) assume the existence of order-reversing action, so that there is no separating equilibrium (signal jamming). In our setting, separating equilibrium exists but it is not optimal under principal's objective.

Our paper is also related to manipulative behaviors in information design problem (e.g. Perez-Richet and Skreta, 2022), in mechanism design problem without monetary transfer (e.g. Perez-Richet and Skreta, 2023), and in classification problem in machine learning setting (e.g. Hardt et al., 2016). Perez-Richet and Skreta (2022) and Perez-Richet and Skreta (2023) focus on fraud-proof mechanisms, i.e., agents have no incentive to exert effort on-path, while our paper allows for agents to exert effort on-path. Hardt et al. (2016) considers a single agent classification problem, while our paper tackles a multi-agent resource allocation problem where agents can manipulate signals. The coordination of efforts among agents is one of the main challenge in our model, which is absent in Hardt et al. (2016).

Our paper is also related to a literature on strategic communication with lying cost. A major departure from this literature is that we study an allocation problem.

2 An Example

Suppose there are two agents and one principal. Each agent i has a private type θ_i drawn from a uniform distribution $F(\theta) = \theta$ with support on $[0, 1]$. Notice that the private type here reflects the hidden ability of the agent, instead of the private valuation of the item.[8] Each agent i can privately choose a non-negative effort e_i to produce a public signal $s_i \in [0, \infty)$. The effort is $e_i = \max\{0, s_i - \theta_i\}$. This specification captures the realistic feature that agents can generate any signal lower than their own type for free. For example, a talented student can easily pretend to be a not so good one by purposely writing down wrong answers in any exam. This assumption also captures the phenomenon of head starts in real world competition (See Siegel, 2014). For instance, university instructors can evaluate students based on their tests scores, and more talented students can achieve higher scores with less effort than less talented students. For simplicity, assume that effort cost equals to effort level.

The principal has one item to allocate between the two agents based on the public signal, i.e., the principal chooses an allocation vector (x_1, x_2) with $0 \leq x_i \leq 1, i \in \{1, 2\}$ and $x_1 + x_2 \leq 1$. Each agent's valuation for the item is 1. His utility is the value of the item he receives subtracts the cost of effort, i.e., $u_i = x_i - e_i$. The *matching efficiency* of an allocation (x_1, x_2) is measured by $\theta_1 \cdot x_1 + \theta_2 \cdot x_2$. Under perfect assortative matching, i.e., the highest type gets the item, matching efficiency is the highest. Principal values both matching efficiency and minimizing agents' effort and values them equally. That is, principal's objective is to maximize $\sum_{i=1}^{2}(\theta_i x_i - e_i)$. To understand this objective function, one should view the principal as a benevolent social planner or some government agencies, who cares about both matching efficiency and all participants' utilities. Matching efficiency captures the efficiency concern to allocate the best resources to the most in-need individuals. For example, in a government subsidy program with limited budget, the firm or individual with most financial need should be given most financial help. In college admissions, it's for the society's interest to match the brightest student to the best education resources. However, competing for limited resources could induce wasteful effort. It is of society's interest to minimize such effort.

Suppose for a second the principal only cares about matching efficiency. A natural candidate for implementing the efficient allocation in classic auction environment is the second-price auction, i.e., the highest agent wins and pays the threshold payment (the second-price). However, since monetary payments are not allowed in our setting, the principal can only use signals as an alternative instrument instead of monetary payments. In order for the winning agent to be indifferent between winning and losing at the threshold type (second highest type), the recommended signal to the winning agent does not equal the second highest type, but equals one (the value for winning the item) plus the second highest type. We call this a "second-price" format mechanism.

[8] One can think of the type as the cost type in the signalling literature, instead of the valuation type as in the mechanism design literature.

Example 1 ("Second-price" format mechanism)**.** *Each agent i reports his own type $\hat{\theta}_i$. For each reported type profile $\hat{\boldsymbol{\theta}}$, let $i^* = \arg\max_i \hat{\theta}_i$ be the agent with the highest reported type, and $s^* = 1 + \max_{i \neq i^*} \hat{\theta}_i$ the prescribed signal that the agent has to produce in order to get the item.*[9]

- *Principal recommends agent i^* to generate signal s^* and another agent $i \neq i^*$ to generate signal 0.*
- *Principal allocates the item to agent i^* if and only if the observed signal for i^* is at least s^*. Otherwise the principal keeps the item.*

Using a similar argument as in second price auction, each agent has incentive to truthfully report their type and produce the signal as prescribed. In this mechanism, the item is allocated to the agent with the highest type. That is, the principal achieves the first best outcome in terms of matching efficiency. However, this mechanism is not desirable in terms of minimizing expected effort because in order to create the incentive for each agent to truthfully report their type, the agent with the highest type has to "burn" some utility as a proof of being the highest type. More specifically, the highest type has to exert strictly positive effort $(1 + \mathbf{E}\big[\theta_{(2)} | \theta_{(2)} \leq \theta_{(1)}\big] - \theta_{(1)}) \in (0, 1)$, where $\theta_{(2)}$ is the second highest type and $\theta_{(1)}$ is the highest type.

Next we propose another mechanism, a winner-takes-all contest, where the principal can improve on her objective value if she also cares about minimizing agents' effort cost. A contest is a special mechanism that allocates the item based *solely* on the rankings of the agents' performance profile which are the observed signals in our setting.[10]

Example 2 (Contest)**.** *Principal commits to a contest rule that allocates the item to the agent who generates the highest signal. Under this game rule, each agent using the strategy $s_i(\theta_i) = \theta_i$ is an equilibrium.*

Given another agent is using the strategy $s_{-i}(\theta_{-i}) = \theta_{-i}$, agent i's payoff of exerting effort e_i is $1 \cdot F(\theta_i + e_i) - e_i = \theta_i$. Each agent has no strict incentive to exert effort and $s_i(\theta_i) = \theta_i$ is a best response. Therefore, each agent producing a signal that equals to his own type is an equilibrium. Notice that this contest also allocates the item to the agent with the highest type and in equilibrium, each agent exerts zero effort. This achieves the upper bound of the principal's objective value. Therefore, this contest is optimal.

[9] Here the number 1 is the agent's valuation of the item. Under this recommendation, agent i's utility of reporting $\hat{\theta}_i$ given his own type θ_i and another agent's reported type $\hat{\theta}_{-i}$ is $\theta_i - \hat{\theta}_{-i}$ if $\hat{\theta}_{-i} < \hat{\theta}_i$ and is 0 otherwise. This gives type θ_i a utility 0 when $\hat{\theta}_{-i} = \hat{\theta}_i$. However, if the recommended signal is lower, then type θ_i has a positive utility when $\hat{\theta}_{-i} = \hat{\theta}_i$, which creates incentive to misreport a higher type. Similar reasoning applies if the recommended signal is higher. Hence, $s^* = 1 + \max_{i \neq i^*} \hat{\theta}_i$ is the only recommendation that works.

[10] Notice that the mechanism in Example 1 is not a contest because the winning agent has to produce a signal that depends on another agent's type.

Moreover, in both the "second-price" format mechanism and the contest, the efficient allocation rule is implemented, and the lowest type gets utility zero since such type receives the item with probability zero by exerting effort zero. However, in the contest, each agent i's expected utility is $F(\theta_i) = \theta_i$, while in the "second-price" format mechanism, each agent's expected utility is $(\theta_i - \mathbf{E}[\theta_{-i}|\theta_{-i} \leq \theta_i])F(\theta_i) < F(\theta_i)$ for any $\theta_i > 0$. These observations immediately imply that payoff-equivalence result fails to hold in our setting.

Unlike in mechanism design settings with monetary transfers, in our setting the alternative instrument is agents' effort. However, effort is not enforceable because it is not observable, and signal is the observable and hence contractible object. In a stochastic mechanism, this introduces the possibility of double deviation, since both type and effort are not observable. For a given prescribed signal, the agent could first lie about his type and later make it up by exerting the right amount of effort.

In the "second-price" format mechanism, the prescribed signal depends on all agents types. From each agent's point of view, the prescribed signal is stochastic because it depends on other agents' type realization. The double deviation strategy becomes attractive under this mechanism, because each agent could misreport as a higher type and later, depending on the realized prescribed signal, the agent could either produce the signal, if it does not require too much effort cost, or opt out, if on the contrary it requires too much effort cost. Put it in a different way, the prescribed signal reveals some information about other agents type, which better guides the agents whether it is profitable to pretend to be a high type. Therefore, the principal has to set a prescribed signal that is high enough to deter such double deviations, which makes implementing the same efficient allocation rule more "expensive" in terms of the effort cost it induces, compared to the scenario where the prescribed signal does not leak out any information about other agents' type. Contests rule out such information leakage. In contests, principal commits to a contest rule, which specifies a Bayesian game for the agents. Each agent hence chooses a strategy that does not depend on other agents' types.[11]

It turns out that these observations are not confined to the current simple setting. In the following sections, we will show that contests being optimal remains to hold in more general settings. In particular, "second-price" format mechanism is strictly dominated by winner-takes-all contest. However, we will reverse the order of exposition in the formal analysis. Here is a sketch of our approach using this simple example. To compare the above two mechanisms more closely, notice that the contest can be implemented by a direct mechanism where each agent i reports his own type $\hat{\theta}_i$. Given each agent's reported type $\hat{\theta}_i$, principal recommends each agent i to generate a signal $s_i(\hat{\theta}_i) = \hat{\theta}_i$. The item is allocated to the agent i that generates the highest signal. The major difference between these two direct mechanisms is that in the contest, the signal recommendation

[11] Viewed as direct mechanisms, contests are deterministic mechanisms that rule out the possibility of double deviations.

is solely based on each agent's own type. It turns out that the reverse is also true: each member in the smaller class of mechanisms where signal recommendation made to each agent could only depend on his own type, indeed has an indirect implementation that is consistent with what we usually call a contest, i.e., a game that allocates the item(s) to the agents based on the ranking of their performance.

In Section 3, we formally define the model. In Section 4, we first define what the most general mechanism could look like, and then a smaller class of mechanisms, where signal recommendation made to each agent could only depend on his own type. We call this smaller class of mechanisms implementable by contests, or contests, without justification. Then we prove that the optimal mechanism falls into the smaller class of mechanisms which we call contests. Only in Section 6, we will demonstrate that the mechanisms we call contests are consistent with the class of games that we usually call contests in the contest literature.

3 Models

The principal (she) allocates k units of identical items to $n > k$ heterogeneous agents (he). The allocation $\boldsymbol{x} = (x_i)_{i=1}^n$ is a vector of probability such that $0 \leq x_i \leq 1$ and $\sum_{i=1}^n x_i \leq k$. Let $X \subseteq [0,1]^n$ be the space of all such (randomized) allocations. Each agent i has a private type θ_i drawn independently from a publicly known distribution F_i supported on $\Theta_i = [\underline{\theta}_i, \bar{\theta}_i] \in \mathbb{R}_+$. Denote the distribution over type profile $\boldsymbol{\theta}$ by $\boldsymbol{F}$. For simplicity, we assume that the density function f_i exists for all i, and $f_i(\theta_i) > 0$ for any $\theta_i \in [\underline{\theta}_i, \bar{\theta}_i]$.

The principal cannot directly observe the private types but can rely on the public signal profiles for allocating the items. Specifically, each agent i can generate a public signal $s_i \in S_i = \mathbb{R}_+$ with effort $e(s_i, \theta_i) = (s_i - \theta_i)^+ := \max\{0, s_i - \theta_i\}$. In other words, each agent's signal is the sum of own type and effort, given that the signal is higher than his type. Unlike the classic setting in the signalling literature, in our model, being a higher type does not make effort more productive in terms of generating signal, i.e., signal is not modelled as the product of effort and type in our setting. We argue that our formulation better captures the applications where gaming effort and innate ability are orthogonal. For instance, in certain school subjects, test scores can be improved by both learning logical reasoning and "learning stuff by heart", and the latter is not made easier or harder by the ability on logical reasoning. In grant applications, more effort may just produce longer or more beautifully written grant applications, which is unrelated to the underlying quality of the work. Moreover, these type of signals are independent of the originality of the research as such (which is what grant bodies care about).[12]

Assume that agent i's cost of effort is $\eta \cdot e(s_i, \theta_i)$, where $\eta > 0$ is the ability of the agent i for

[12]We thank Carl-Christian Groh for offering us these examples.

manipulating the signals, or the marginal cost of effort and it is publicly known.[13] Our main results do not rely on the linear structure of the cost function. However, linearity simplifies the analysis and the exposition. In Appendix B, we extend our results under convex cost structure.

Payoffs. Each agent has unit-demand for the items. The utility of agent i for getting a single unit of the items with probability $x_i \in [0,1]$ when generating signal s_i is [14]

$$u_i = x_i - \eta \cdot e(s_i, \theta_i).$$

The *matching efficiency* of an allocation $\boldsymbol{x}$ is measured by $\sum_i \theta_i \cdot x_i$. Under perfect assortative matching, i.e., the highest k types get one unit of item, matching efficiency is the highest. The principal's payoff for choosing an allocation $\boldsymbol{x}$ and utility profile $\boldsymbol{u} = (u_i)_{i=1}^n$ is

$$\sum_{i=1}^{n} \alpha \cdot \theta_i \cdot x_i + (1-\alpha) \cdot u_i, \tag{1}$$

for some $\alpha \in [0,1]$. To understand this objective function, one should view the principal as a benevolent social planner or some government agencies, who cares about both matching efficiency and all participants' utilities. To fix ideas, consider college admissions. A social planner would like to match the brightest students and the best universities, because it increases the future productivity of the students. Hence, the matching efficiency appeared in the objective function is a reduced form approach to capture the long-term benefit of allocating the best resources in the society to the most capable individuals. However, aware that the selection criteria can be influenced by wasteful effort, a social planner who cares about participants' welfare would like to minimize such effort.

General Mechanism Design. We restrict our attentions to direct mechanisms. The principal can communicate with all the agents before their choice of effort and make signal recommendations based on the communication. By the revelation principle, it is without loss to focus on direct mechanisms where the agents first report their private types to the principal, and then the principal recommends the signals to the agents based on the aggregated reports. Formally, the timeline for a

[13] It is without loss to assume that η is publicly known. If η is private, the principal can easily elicit η by asking each agent to report η in addition to the type and never allocates to those who report differently from others. Truthfully reporting η is an equilibrium.

[14] Notice that each agent's utility does not satisfy single-crossing in our setting. For any type $\theta \in \Theta_i$, if he prefers (x_1, s_1) to (x_2, s_2), then either one of the following is true: (1)$x_1 \geq x_2$ if $\min\{s_1, s_2\} \leq \theta$; (2) $x_1 - s_1 \geq x_2 - s_2$ if $\min\{s_1, s_2\} \geq \theta$; (3) $x_1 - s_1 + \theta \geq x_2$ if $s_2 < \theta \leq s_1$; (4) $x_1 \geq x_2 - s_2 + \theta$ if $s_1 < \theta \leq s_2$. It is easy to see from the last case that single crossing does not hold. This contrasts with classic mechanism design and signalling game setting where strict single crossing is usually assumed. In the gaming literature (e.g., Kartik, 2009; Ball, 2019), agent's utility does not satisfy single crossing but these papers make different assumption on agent's utility.

general mechanism is described as follows.

1. The principal commits to a signal recommendation policy $\tilde{s} : \prod_{i=1}^{n} \Theta_i \to \Delta(S)$ and allocation rule $p : \prod_{i=1}^{n} \Theta_i \times S \to X$.

2. Each agent i reports type θ_i' to the principal and receives signal recommendation $\tilde{s}_i(\boldsymbol{\theta}')$. Then each agent i chooses signal $s_i' \in \{\tilde{s}_i(\boldsymbol{\theta}'), 0\}$.[15]

3. Principal observes signal profile $\boldsymbol{s}'$ and each agent i receives the item with probability $p_i(\boldsymbol{\theta}', \boldsymbol{s}')$.

Define the interim allocation rule as $Q_i(\theta_i) = \mathbf{E}_{\boldsymbol{\theta}_{-i}}\left[\mathbf{E}_{s_i \sim \tilde{s}_i(\theta_i, \boldsymbol{\theta}_{-i})}[p_i(\theta_i, \boldsymbol{\theta}_{-i}, s_i, \boldsymbol{s}_{-i}(\boldsymbol{\theta}_{-i}))]\right]$, and the interim utility of agent i as $U_i(\theta_i) = \mathbf{E}_{\boldsymbol{\theta}_{-i}}\left[\mathbf{E}_{s_i \sim \tilde{s}_i(\theta_i, \boldsymbol{\theta}_{-i})}[p_i(\theta_i, \boldsymbol{\theta}_{-i}, s_i, \boldsymbol{s}_{-i}(\boldsymbol{\theta}_{-i})) - \eta \cdot e(s_i, \theta_i)]\right]$.

The class of general mechanisms defined here is slightly larger than the class of mechanisms considered in the classic setting (e.g. Myerson, 1981), where allocation is determined by the reported type profile. In our setting, monetary transfer is not available and effort recommendation, equivalently, signal recommendation, is the additional instrument to discipline agents' incentives. In any mechanism that assigns the items based on the reported type profile but not the actual signals, agents' reports are purely cheap talk and agents have strong incentives to misreport. Since effort is unobservable, each agent could benefit from reporting a type as high as possible to secure the allocation, later produce a signal that is no higher than his true type to save effort cost. Hence allowing the allocation rule to depend on both reported type profile and the actual signals is a more natural consideration.

We provide the formal definition of any interim allocation-utility pair that $(\boldsymbol{Q}, \boldsymbol{U})$ is *implementable by a general mechanism.*

Definition 1. *An interim allocation-utility pair* $(\boldsymbol{Q}, \boldsymbol{U})$ *is* ***implementable by a general mechanism*** *if*

1. $(\boldsymbol{Q}, \boldsymbol{U})$ *is induced by some signal recommendation policy* $\tilde{\boldsymbol{s}}$ *and an allocation rule* $\boldsymbol{p}$*:*

$$\begin{aligned} Q_i(\theta_i) &= \mathbf{E}_{\boldsymbol{\theta}_{-i}}\left[\mathbf{E}_{s_i \sim \tilde{s}_i(\theta_i, \boldsymbol{\theta}_{-i})}[p_i(\theta_i, \boldsymbol{\theta}_{-i}, s_i, \boldsymbol{s}_{-i}(\boldsymbol{\theta}_{-i}))]\right] \\ U_i(\theta_i) &= Q_i(\theta_i) - \eta \cdot \mathbf{E}_{\boldsymbol{\theta}_{-i}}\left[\mathbf{E}_{s_i \sim \tilde{s}_i(\theta_i, \boldsymbol{\theta}_{-i})}[e(s_i, \theta_i)]\right], \forall i, \theta_i \end{aligned} \qquad \text{(Consistency)}$$

2. *each agent has weak incentive to truthfully report his own type and follow the signal recommendation, i.e.,*

$$U_i(\theta_i) \geq \mathbf{E}_{\boldsymbol{\theta}_{-i}}\left[\mathbf{E}_{s_i \sim \tilde{\boldsymbol{s}}(\theta_i', \boldsymbol{\theta}_{-i})}\left[\max\{0, p_i(\theta_i', \boldsymbol{\theta}_{-i}, s_i, \tilde{s}_{-i}(\theta_i', \boldsymbol{\theta}_{-i})) - \eta \cdot e(s_i, \theta_i)\}\right]\right], \quad \forall i, \theta_i, \theta_i'. \qquad \text{(Incentives)}$$

[15] This is without loss of generality. The principal can partially enforce the recommendation by assigning no item to any agent who chooses a signal different from the recommended one. Hence each agent essentially has two choices: following the recommendation ($s_i' = \tilde{s}_i(\boldsymbol{\theta}')$) and opting out ($s_i' = 0$).

We say an allocation rule $\boldsymbol{Q}$ is **implementable by a general mechanism** if there exists an interim utility profile $\boldsymbol{U}$ such that $(\boldsymbol{Q}, \boldsymbol{U})$ is implementable by a general mechanism.

In general mechanisms (as opposed to contests defined in the following paragraphs), principal can communicate with all the agents and the signal recommendation for each agent can potentially depend on the reported type profile.

Next, we introduce a smaller class of mechanisms that are *implementable by contests*.

Contest Design. The principal commits to a contest rule $\boldsymbol{x} : S \to X$, which is a mapping from the realized signal profiles to the (randomized) allocation. There is no communication between the principal and the agents before their effort choices. Each agent chooses a signal strategy after the announcement of the contest rule. It is worth emphasized that the effort choice of each agent is not correlated with the realized types or effort choices of other agents.

We define the strategy of each agent i as $s_i : \Theta_i \to \Delta(S_i)$. Then the item is distributed according to the realized signal profile $\boldsymbol{s} = (s_i)_{i=1}^n$ given the contest rule $\boldsymbol{x}(\boldsymbol{s})$. The timeline of the contest model is formalized as follows.

1. Principal commits to a contest rule $\boldsymbol{x} : S \to X$.
2. Each agent i with type θ_i generates a signal $s_i(\theta_i)$ with cost $e_i(s_i(\theta_i), \theta_i)$.
3. Principal observes the signal profile s and each agent i receives the item with probability $x_i(\boldsymbol{s})$.

Given a contest rule $\boldsymbol{x}$, it defines a game, where agents' equilibrium strategy profile is $\boldsymbol{s} = (s_i)_{i=1}^n$. Given a strategy profile $\boldsymbol{s}'$, agent i's payoff is $v_i(\boldsymbol{s}', \theta_i) = x_i(\boldsymbol{s}') - \eta \cdot e_i(s_i', \theta_i)$. In equilibrium, agent i's expected utility is $u_i(s_i; \theta_i) = \mathbf{E}_{\boldsymbol{\theta}_{-i}}[v_i(s_i, \boldsymbol{s}_{-i}(\boldsymbol{\theta}_{-i}), \theta_i)]$.

The equilibrium analysis approach is notoriously intractable (See for instance Olszewski and Siegel, 2016). Instead of adopting this approach, we utilize mechanism design technique to analyze contests. This is possible thanks to the useful observation that a contest can be viewed as a direct mechanism where agents report their types to the principal, and receives a signal recommendation *solely* based on his own reported type. In the direct mechanism that implements the equilibrium $\boldsymbol{s}$ induced by the contest x, we can define the ex post allocation rule as $q_i(\theta_i, \boldsymbol{\theta}_{-i}) = x_i(s_i(\theta_i), \boldsymbol{s}_{-i}(\boldsymbol{\theta}_{-i}))$, the interim allocation rule as $Q_i(\theta_i) = \mathbf{E}_{\theta_{-i}}[q_i(\theta_i, \theta_{-i})]$, and the interim utility as $U_i(\theta_i) = u_i(s_i(\theta_i); \theta_i)$. Let an interim allocation rule be $\boldsymbol{Q} = (Q_i)_{i=1}^n$, and $\boldsymbol{U} = (U_i)_{i=1}^n$ be the profile of interim utilities.

Definition 2. *An interim allocation-utility pair $(\boldsymbol{Q}, \boldsymbol{U})$ is **implementable by a contest** if there exists a contest rule $\boldsymbol{x}$ and a signal strategy $\hat{\boldsymbol{s}}$ such that*

1. $(\boldsymbol{Q}, \boldsymbol{U})$ *is induced by a contest rule* $\boldsymbol{x}$ *and signal strategies* $\hat{\boldsymbol{s}}$*, where* $\hat{s}_i : \Theta_i \to S, \forall i$*:*

$$\begin{aligned} Q_i(\theta_i) &= \mathbf{E}_{\theta_{-i}}[x_i(\hat{s}_i(\theta_i), \hat{s}_{-i}(\theta_{-i}))] && \text{(Consistency)} \\ U_i(\theta_i) &= Q_i(\theta_i) - \eta \cdot e(\hat{s}_i(\theta_i), \theta_i), \quad \forall i, \theta_i \end{aligned}$$

2. *signal strategies* $\hat{\boldsymbol{s}}$ *form an equilibrium, i.e.,*

$$U_i(\theta_i) \geq \mathbf{E}_{\theta_{-i}}\left[x_i(s_i', \hat{s}_{-i}(\theta_{-i}))\right] - \eta \cdot e(s_i', \theta_i), \quad \forall i, s_i', \theta_i. \qquad \text{(Incentives)}$$

We say an allocation rule $\boldsymbol{Q}$ is **implementable by a contest** if there exists an interim utility profile $\boldsymbol{U}$ such that $(\boldsymbol{Q}, \boldsymbol{U})$ is implementable by a contest.

In our definition, we do not impose the requirement that the contest rule $\boldsymbol{x}$ is a mapping from ranking of the signal to allocation. In Section 6, we will show that by generalizing the commonly used notion *strict ranking* to our notion of ranking called *coarse ranking*, such a restriction is redundant.

As we have alluded earlier, the main difference between implementability by a contest and by a general mechanism is that in a contest, the recommended signal for agent i is a function of agent i's *own reported type*, while in a general mechanism, the recommended signal for agent i could be a function of the *whole reported type profile*. This difference in definition per se does not explain why the more restricted class of mechanisms is called contests. However, as we will show in Section 6, by restricting recommended signals to depend on each agent's own type, it allows us to construct the contest rule, a mapping from the ranking of each agent's performance to the allocation, that implements the mechanism.

Interim Approach In the rest of the paper, instead of analyzing the set of ex-post allocation rules, i.e., mappings from type profile to allocation, we perform our analysis on the set of interim allocation rules, i.e., mappings from agent's own type to allocation. Instead of analyzing the two instruments interim allocation rule and signal recommendations, we analyze interim allocation rule and interim utilities. However, not all interim allocation rules are *feasible*. Let an interim allocation rule $\boldsymbol{Q} = (Q_i)_{i=1}^n$ be a profile of mappings with $Q_i : \Theta_i \to [0, 1]$. Let an ex-post allocation rule $\boldsymbol{q} = (q_i)_{i=1}^n$ be a profile of mappings with $q_i : \prod_{i=1}^n \Theta_i \to X$. We say the interim allocation rule $\boldsymbol{Q}$ is *interim feasible* if there exists an ex post allocation rule $\boldsymbol{q}$ such that $Q_i(\theta_i) = \mathbf{E}_{\theta_{-i}}[q_i(\theta_i, \theta_{-i})]$ for any agent i and type θ_i.

With these renamings and technical considerations, the design problem is equivalent to a design problem where the principal chooses a feasible interim allocation rule $\boldsymbol{Q}$ and interim utility profile

$\boldsymbol{U}$ to maximize

$$\mathrm{Obj}_\alpha(\boldsymbol{Q},\boldsymbol{U}) = \mathbf{E}_{\boldsymbol{\theta}}\left[\alpha\cdot\sum_i \theta_i\cdot Q_i(\theta_i) + (1-\alpha)\cdot\sum_i U_i(\theta_i)\right]. \tag{2}$$

4 Optimality of Contests

The main goal of this section is to show that contests are optimal. In order to state our result, we first provide several lemmas in terms of characterizing incentive compatibility constraints in contests and general mechanisms, which turn out to be useful in establishing the main result.

IC characterization Next, we characterize the incentive compatibility conditions in any direct mechanism that implements monotone allocation rule. In general, there exist non-monotone interim allocations that are implementable by contests and by general mechanisms (see examples in Section 8.1). However, as we will show in Theorem 1, this is without loss of optimality.

Lemma 1. *An interim allocation-utility pair* $(\boldsymbol{Q},\boldsymbol{U})$ *with monotone* $\boldsymbol{Q}$ *is* implementable by a contest *if and only if* $\boldsymbol{Q}$ *is interim feasible, and for any agent* i *with type* θ_i,[16]

$$(1)\ U_i'(\theta_i)\in[0,\eta]; \qquad (2)\ U_i(\theta_i)\le Q_i(\theta_i); \qquad (3)\ U_i'(\theta_i)=\eta \text{ if } U_i(\theta_i)<Q_i(\theta_i). \tag{IC}$$

For any interim allocation-utility pair $(\boldsymbol{Q},\boldsymbol{U})$ that is implementable by a contest, there is a level constraint and a slope constraint on interim utility. The level constraint is intuitive: due to the non-negativeness of the effort costs, the utility of the agent is upper bounded by his allocation. The slope constraint says that marginal increase of the interim utility is upper bounded by the marginal cost of effort, otherwise low types would have incentives to misreport his type as a higher type and produce the recommended signal for the higher types. Finally, if the level constraint is slack at any type θ, the equilibrium effort for type θ must be strictly positive. In order to eliminate higher types' incentives of deviating to θ, the slope constraint must be binding at type θ. The proof of Lemma 1 is provided in Appendix A.1. Similarly, we can derive a necessary condition for incentive compatibility condition in a general mechanism. The proof is also provided in Appendix A.1.

Lemma 2. *An interim allocation-utility pair* $(\boldsymbol{Q},\boldsymbol{U})$ *is* implementable by a general mechanism *only if* $\boldsymbol{Q}$ *is interim feasible, and for any agent* i *with type* θ_i,

1. $U_i'(\theta_i)\in[0,\eta]$;

[16] The function U_i may not be differentiable everywhere. For any type θ_i such that U_i' is not differentiable at θ_i, we use $U_i'(\theta_i)$ to denote any subgradient (or simply, left and right derivative) of function U_i'. It is not hard to show that U_i is a monotone function and hence it is differentiable almost everywhere.

2. $U_i(\theta_i) \leq Q_i(\theta_i)$.

The characterization of IC in general mechanisms are more challenging. This is because the principal can make a signal recommendation that depends on all agents' reports, each agent's effort recommendation in general mechanisms is essentially stochastic, while in the mechanism design literature, it is typically without loss to assume deterministic mechanisms.

Payoff Equivalence From Lemma 1, we can establish Proposition 1, which says that for any implementable allocation and utility pair, the utility function for each agent is uniquely pinned down by the interim allocation, up to the choice of the utility for the lowest type.

Proposition 1. *Fix any monotone and interim feasible allocation rule* $\boldsymbol{Q}$*, and any* $\{\underline{u}_i\}_{i=1,\ldots,n}$ *such that* $\underline{u}_i \leq Q_i(\underline{\theta}_i)$ *for all* i*. There exists a unique interim utility profile* $\boldsymbol{U}$ *with* $U_i(\underline{\theta}_i) = \underline{u}_i$ *for all* i *such that* $(\boldsymbol{Q}, \boldsymbol{U})$ *is implementable by a contest. Moreover, for any interim allocation-utility pair* $(\boldsymbol{Q}, \boldsymbol{U}^\dagger)$ *that is implementable by a contest, we have that*

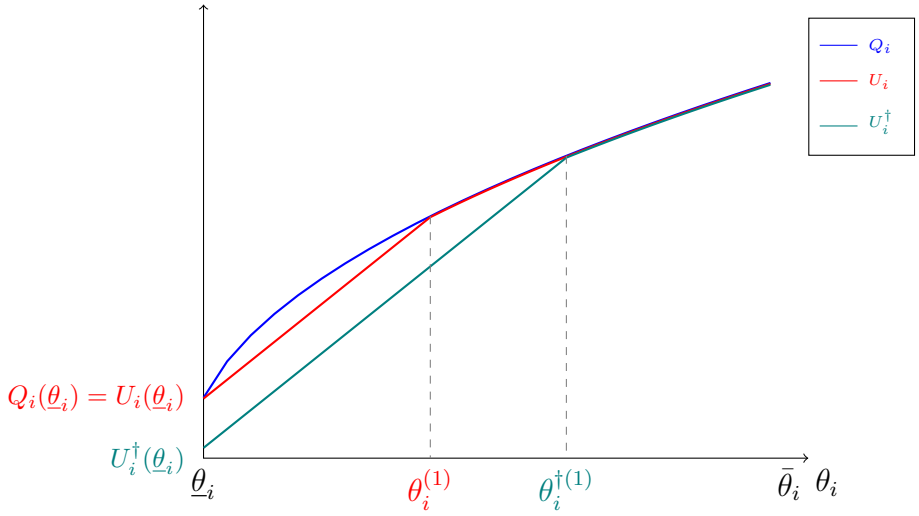
- *if* $U_i(\underline{\theta}_i) > U_i^\dagger(\underline{\theta}_i)$ *for any agent* i*, then* $U_i(\theta_i) \geq U_i^\dagger(\theta_i)$ *for all type* θ_i *and all* i*;*
- *if* $U_i(\underline{\theta}_i) = U_i^\dagger(\underline{\theta}_i)$ *for any agent* i*, then* $U_i(\theta_i) = U_i^\dagger(\theta_i)$ *for all type* θ_i *and all* i*.*

In the classic mechanism design setting, payoff equivalence means that once the allocation is determined, the curvature of the utility function is fixed, and can only be shifted by a constant determined by the utility of the lowest type. In our setting, fixing the allocation rule, shifting the utility of the lowest type does not shift the utilities for all types by the same constant (See Figure 1). We illustrate the construction in Figure 1. For any agent i, if the utility of the lowest type is lower than the interim allocation of the lowest type or the derivative of the interim allocation is larger than parameter η, (IC) imply that the interim utility U_i must be a straight line with derivative η until U_i intersects with Q_i (the intersection type is $\theta_i^{(1)}$ in the example in Figure 1). Then U_i coincides with Q_i until the derivative of Q_i exceeds η. In a discrete type setting, such a recursive procedure could uniquely pin down the interim utility for all types. Unfortunately, the recursive argument fails to work in the continuous type setting. In Appendix A.1, we provide a formal proof to circumvent this technicality.

It is immediate from Proposition 1 that the lowest type exerts zero effort in the optimal contest, because by setting the utility of the lowest type to equal the interim allocation, the utilities of all higher types are weakly improved.

Optimality of Contests We first state a corollary that will be useful to the proof of our main theorem.

Figure 1: Illustrative graph for Proposition 1.



†Both U_i and $U_i^\dagger$ implement the allocation rule Q_i. However, U_i is the implementation that gives agent i the highest utility and hence the best implementation from the principal's point of view. Moreover, U_i and $U_i^\dagger$ do not differ by the same constant that equals $U_i(\underline{\theta}_i) - U_i^\dagger(\underline{\theta}_i)$ as in the standard payoff equivalence result. However, by construction provided in the proof of Proposition 1, U_i is uniquely identified by the allocation rule and $U_i(\underline{\theta}_i)$.

Corollary 1. *Fix any monotone and interim feasible allocation rule $\boldsymbol{Q}$. There exists a unique interim utility profile $\hat{\boldsymbol{U}}$ with $\hat{U}_i(\underline{\theta}_i) = Q_i(\underline{\theta}_i)$ for all i such that $(\boldsymbol{Q}, \hat{\boldsymbol{U}})$ is implementable by a contest. Moreover, $(\boldsymbol{Q}, \hat{\boldsymbol{U}})$ achieves the highest utility for each agent among all general mechanisms that implement $\boldsymbol{Q}$.*

Proof. The first part is directly implied by Proposition 1.

From Lemma 1, we know that for any i, θ_i, if $\hat{U}_i(\theta_i) < Q_i(\theta_i)$, then $\hat{U}_i'(\theta_i) = \eta$. And the contrapositive says if $\hat{U}_i'(\theta_i) < \eta$, then $\hat{U}_i(\theta_i) = Q_i(\theta_i)$. Therefore, $\hat{U}_i(\theta_i)$ can be viewed as a function that is pointwise maximized subject to $\hat{U}_i'(\theta_i) \in [0, \eta]$ and $0 \leq \hat{U}_i(\theta_i) \leq Q_i(\theta_i)$, i.e.,

$$\begin{aligned} \hat{U}_i(\theta_i) = &\max U_i(\theta_i) \\ &\text{s.t. } U_i'(\theta_i) \in [0, \eta] \\ &\quad 0 \leq U_i(\theta_i) \leq Q_i(\theta_i). \end{aligned} \tag{3}$$

Combining with Lemma 2, we know that for any $(\boldsymbol{Q}, \boldsymbol{U})$ that is implementable by a general mechanism, $U_i'(\theta_i) \in [0, \eta]$ and $0 \leq U_i(\theta_i) \leq Q_i(\theta_i)$ must hold. Hence, $\hat{U}_i(\theta_i) \geq U_i(\theta_i)$ for any i and θ_i. □

Now we are ready to state our first main result.

Theorem 1 (Optimality of Contests)**.** *For any interim allocation-utility pair $(\boldsymbol{Q}, \boldsymbol{U})$ that is implementable by a general mechanism, there exists another interim allocation-utility pair $(\boldsymbol{Q}^\dagger, \boldsymbol{U}^\dagger)$ with*

monotone allocation $\boldsymbol{Q}^\dagger$ that is implementable by a contest and attains weakly higher objective value for all $\alpha \in [0,1]$.

Proof. **Part 1**: Lemma 2 implies that $\boldsymbol{Q}$ is interim feasible and $U_i(\theta_i) \leq Q_i(\theta_i)$ for any agent i with type θ_i. If $\boldsymbol{Q}$ is monotone, set $\boldsymbol{Q}^\dagger = \boldsymbol{Q}$ and $\boldsymbol{Q}^\dagger$ is interim feasible. Otherwise, let $G_i(z) = \int_{\underline{\theta}_i}^{\bar{\theta}_i} \mathbf{1}\left[Q(\theta_i) \leq z\right] \mathrm{d}F_i(\theta_i)$ and let $Q_i^\dagger(\theta_i) = G_i^{-1}(F_i(\theta_i))$.[17] Essentially, $Q_i^\dagger$ is a rearrange of allocation Q_i such that $Q_i^\dagger$ is monotone, and the measure of types with allocation at most z are the same for all z.

Using results in Border (1991) and Che et al. (2013), which is restated below, it is easy to verify that $Q^\dagger$ is interim feasible.

Lemma 3 (Che et al., 2013)**.** *Let $w(\boldsymbol{\theta}, \boldsymbol{A}) = |\{i : \theta_i \in A_i\}|$ be the number of agents whose type θ_i is in A_i,*[18] *specified by the given set $\boldsymbol{A} = \prod_{i=1}^n A_i \subset \prod_{i=1}^n \Theta_i$. The interim allocation rule $\boldsymbol{Q}$ is* interim feasible *if and only if*

$$\sum_i \int_{A_i} Q_i(\theta_i) \,\mathrm{d}F_i(\theta_i) \leq \int_A \min\left\{k, w(\boldsymbol{\theta}, \boldsymbol{A})\right\} \mathrm{d}F(\theta) \qquad \forall \boldsymbol{A} = \prod_{i=1}^n A_i \subset \prod_{i=1}^n \Theta_i. \tag{IF}$$

Moreover, for monotone allocations in symmetric environments, (IF) *is equivalent to* [19]

$$\int_\theta^{\bar{\theta}} Q(z) \,\mathrm{d}F(z) \leq \int_\theta^{\bar{\theta}} Q_{\mathrm{E}}(z) \,\mathrm{d}F(z), \quad \forall \theta \in [\underline{\theta}, \bar{\theta}]. \tag{$\widehat{\text{IF}}$}$$

where $Q_{\mathrm{E}}(\theta) = \sum_{j=0}^{k-1} \binom{n-1}{j} \cdot (1 - F(\theta))^j \cdot F^{n-1-j}(\theta)$ is the interim allocation rule for allocating k items efficiently.

As $Q_i^\dagger$ only shifts high allocation from low types to high types, for any agent i,

$$\mathbf{E}_{\theta_i}\left[\theta_i \cdot Q_i^\dagger(\theta_i)\right] \geq \mathbf{E}_{\theta_i}[\theta_i \cdot Q_i(\theta_i)]\,.$$

Part 2: we show that $U_i(\theta_i) \leq Q_i^\dagger(\theta_i)$ is satisfied for all type θ_i. Suppose otherwise, there exist a type θ_i such that $U_i(\theta_i) > Q_i^\dagger(\theta_i)$. As $Q_i^\dagger$ is simply a rearrange of allocation Q_i, there exists a type $\theta_i' \geq \theta_i$ such that $Q_i(\theta_i') \leq Q_i^\dagger(\theta_i)$. By incentive compatibility, $U_i(\theta_i) \leq U_i(\theta_i')$. These three inequalities imply that

$$Q_i(\theta_i') \leq Q_i^\dagger(\theta_i) < U_i(\theta_i) \leq U_i(\theta_i'),$$

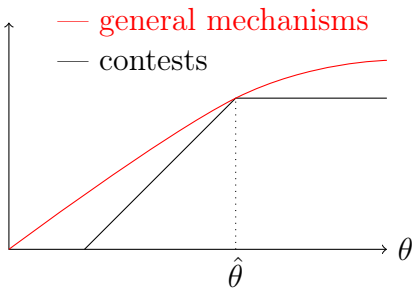
which is a contradiction.

[17] $\mathbf{1}[\cdot]$ is the indicator function.

[18] $|\cdot|$ is the cardinality of the set.

[19] In symmetric environments, by slightly abusing notations, we use $F = F_i$ for all i to denote each agent's type distribution.

Figure 2: Graphic illustration on the intuition of Theorem 1



†The black (red) curve is the utility of each type θ when he misreports his type to be $\hat{\theta}$ in a contest (general mechanism). A general mechanism gives each type higher deviation utility because the signal recommendation reveals information on other agents' types. Hence an agent can decide whether or not to follow the signal recommendation after seeing it. A general mechanism needs to deter this double deviation incentive and hence is more "expensive" to implement.

Part 3: given interim feasible and monotone allocation $\boldsymbol{Q}^\dagger$, let $U^\dagger$ be the utility function stated in Corollary 1. Then $(\boldsymbol{Q}^\dagger, \boldsymbol{U}^\dagger)$ is implementable by a contest. Moreover, combining **Part 2** and $U_i'(\theta_i) \in [0, \eta]$ for all i and θ_i from Lemma 2, we have $U_i^\dagger(\theta_i) \geq U_i(\theta_i)$ for all i and θ_i.

Since both the expected matching efficiency and the agents' expected utilities are weakly improved, the objective value of $(\boldsymbol{Q}^\dagger, \boldsymbol{U}^\dagger)$ is weakly improved compared to $(\boldsymbol{Q}, \boldsymbol{U})$. □

There are two implications of this result. First, for any non-monotone allocation rule that is implementable by a general mechanism, there exists a monotone modification that is both feasible and implementable. Second, for any monotone allocation rule that is implementable by a general mechanism, it is implementable by a contest that can achieve weakly higher value for any weighted average of matching efficiency and agents' utilities. Together, these imply that it is without loss of optimality to restrict out attention to contests with monotone allocations.

The contest structure can be viewed as an all-pay format mechanism, where each agent chooses an effort to compete for the items, regardless of the effort choices of other agents. Our result indicates that this all-pay format is optimal for any choice of α in the objective function, which may seem surprising at first glance as the goal is to minimize costly effort while contests require every agent to pay the cost depending their choice of effort. The intuition is better illustrated in the discrete type setting, which we summarized in Figure 2. By enforcing the all-pay format, we can create the least incentives for agents to deviate from the recommendation of the principal. This helps alleviate the moral hazard issue, which reduces the expected amount of effort we require from each agent to prove that he has the claimed type.

Moreover, the linearity assumption is not crucial for Theorem 1 to hold. However, it ensures that local IC is enough to guarantee global IC and enables us to characterize the IC in closed form, which greatly simplifies the analysis. In Appendix B, we show that contests remain to be optimal for a broad family of convex cost functions.

Greenwald et al. (2018); Zhang (2023) show that the optimal mechanism can be implemented by contests in the setting where effort is productive and principal wants to maximize effort. Our paper differs from their results in at least two aspects. First, effort is unproductive in our setting and the principal aims at maximizing the weighted average of matching efficiency and agents' utilities and efforts are wasteful in our model, while in Greenwald et al. (2018); Zhang (2023), the principal maximizes total efforts. Second, the optimality of contests in previous papers relies on the payoff equivalence result to hold for the class of general mechanisms, while in our model, payoff equivalence fails for all general mechanisms. For instance, the mechanism in Example 1 (Section 2) cannot be implemented by a contest.

A related observation has been made in Perez-Richet and Skreta (2023) Appendix C. They restrict attention to score-based allocation rule that depends only on each agent's score. They argue that this is without loss of generality because (1) their setting can be viewed as a single agent problem due to the continuum of agents; and (2) actions are *contractible* in their setting. In contrast, our model is a multi-agent problem, where principal can potentially benefit from coordinating agents' actions by communication. Therefore, our problem is more involved and our result is non-trivial.

Sub-optimality of "Second-price" format mechanism

Definition 3 ("Second-price" format mechanism for agents with unit demand)**.** *A mechanism is of "second-price" format if it admits the following form:*

- *each agent reports his own type $\hat{\theta}_i$;*
- *for each reported type profile $\hat{\boldsymbol{\theta}}$, let $I^* = \{i : \hat{\theta}_i$ is the k highest$\}$ be the set of agents with the k highest reported types, and $s^*_{i^*} = \frac{1}{\eta_{i^*}} + \max_{i \notin I^*} \hat{\theta}_i$ the prescribed signal that agent $i^* \in I^*$ has to produce in order to get one unit of item;*
- *Principal recommends agent i^* to generate signal $s^*_{i^*}$ and other agents $i \neq i^*$ to generate signal 0.*
- *Principal allocates one unit of item to agent i^* if and only if the observed signal for i^* is at least $s^*_{i^*}$. Otherwise the principal keeps the item.*

Definition 4 (WTA contest for agents with unit demand)**.** *If the agent produces one of the k highest signals, then he gets one unit of the item.*

Proposition 2. *For any $\alpha < 1$, any $n \geq 2$, any $k < n$ and for all except at most one distrbituion F, "second-price" format mechanism (for agents with unit demand) is strictly dominated by WTA contest (for agents with unit demand).*

5 Optimal Contests

Based on Theorem 1, principal's designed problem can be simplified as choosing a monotone and feasible interim allocation rule and an interim utility profile such that the interim allocation-utility pair is implementable by contest which is equivalent to imposing the set of constraints described in (IC). That is, principal solves

$$\begin{aligned} V_\alpha = \sup_{\boldsymbol{Q},\boldsymbol{U}} \quad & \mathrm{Obj}_\alpha(\boldsymbol{Q},\boldsymbol{U}) \\ \text{s.t.} \quad & \boldsymbol{Q} \text{ is monotone, interim feasible,} \\ & (\boldsymbol{Q},\boldsymbol{U}) \text{ satisfies (IC)} \end{aligned} \tag{$\mathcal{P}_\alpha$}$$

From now on, we use the term contest and the term direct mechanism that is implementable by contest interchangeably when it does not cause confusion. Hence the task of finding the optimal mechanism is essentially the task of finding the optimal contest. We also assume that agents are ex-ante homogeneous in the rest of the paper to simplify the analysis.[20]

Assumption 1 (Symmetric Environments)**.** *Agents are ex-ante homogeneous, i.e.,* $\Theta_i = \Theta = [\underline{\theta}, \bar{\theta}]$, $F_i = F$ *and* $f_i = f$ *for all* i.

Note that for Program ($\mathcal{P}_\alpha$), although the objective function is linear, the (IC) constraints are not convex. This is, a convex combination of two allocation-utility pairs may violate the (IC) constraints. Nonetheless, in the following lemma, we show that the optimal contest for this non-convex optimization program is always symmetric in symmetric environments.

Lemma 4. *Under Assumption 1, the optimal contest is symmetric, for any* $\alpha \in [0,1]$.

Proof. Consider a relaxed problem $\mathcal{P}'_\alpha$ where instead of the (IC) constraints, we only require that $U_i'(\theta_i) \in [0,\eta]$ and $U_i(\theta_i) \le Q_i(\theta_i)$ hold for any agent i with type θ_i. Note that this is a convex constraint, and hence the relaxed problem is a convex program. Thus, there exists a symmetric optimal solution $(\boldsymbol{Q},\boldsymbol{U})$ for program $\mathcal{P}'_\alpha$ if the environment is symmetric. Moreover, as $\boldsymbol{U}$ is maximized given the derivatives constraint and the upper bound of $\boldsymbol{Q}$, the allocation-utility pair $(\boldsymbol{Q},\boldsymbol{U})$ also satisfies the (IC) constraints according to the proof of Proposition 1. Therefore, $(\boldsymbol{Q},\boldsymbol{U})$ is also feasible and hence an optimal solution for Program ($\mathcal{P}_\alpha$). □

By restricting attention to symmetric contests, the design of optimal contests reduces to the optimization for the single agent problem for any particular agent i. We omit the subscript i from the notation for this single agent problem when there is no ambiguity. When there is no confusion, we use the interim allocation rule for each agent Q to refer to the interim allocation profile, and each

[20] See Section 8.4 for discussion under asymmetric environment.

agent's utility function U to refer to the interim utility profile in the symmetric environment. Let $Q_{\mathrm{E}}(\theta)$ be the interim allocation rule for maximizing matching efficiency, i.e., the efficient allocation rule. The optimization program can be reformulated as follows.

$$\begin{aligned} \hat{V}_\alpha = \sup_{Q,U} \quad & \mathbf{E}_\theta[\alpha \cdot \theta \cdot Q(\theta) + (1-\alpha) \cdot U(\theta)] \\ \text{s.t.} \quad & Q \text{ is monotone,} \\ & \int_\theta^{\bar{\theta}} Q(z)\,\mathrm{d}F(z) \le \int_\theta^{\bar{\theta}} Q_{\mathrm{E}}(z)\,\mathrm{d}F(z), \quad \forall \theta \in [\underline{\theta}, \bar{\theta}] \\ & (Q,U) \text{ satisfies (IC)} \end{aligned} \tag{$\hat{\mathcal{P}}_\alpha$}$$

The simplification of interim feasibility for symmetric environments is given by Che et al. (2013).

Let (Q_α, U_α) be the optimal solution for program $(\mathcal{P}_\alpha)$.[21] The following theorem implies that the optimal allocation partitions the type space into three types of intervals.

Theorem 2. *Under Assumption 1, for any $\alpha \in (0,1)$, the optimal contest (Q_α, U_α) defines an interval partition $\{(\underline{\theta}^{(j)}, \bar{\theta}^{(j)})\}_{j=1}^{\infty}$ of the type space.*[22] *For any $j \ge 1$, the interval $(\underline{\theta}^{(j)}, \bar{\theta}^{(j)})$ belongs to exactly one of the following three regions:*[23]

1. no-tension region*, where $Q_\alpha(\theta) = U_\alpha(\theta) = Q_{\mathrm{E}}(\theta)$ and $U'_\alpha(\theta) < \eta$ for any type $\theta \in (\underline{\theta}^{(j)}, \bar{\theta}^{(j)})$;*

2. no-effort region*, where $Q_\alpha(\theta) = U_\alpha(\theta)$ and $U'_\alpha(\theta) = \eta$ for any type $\theta \in (\underline{\theta}^{(j)}, \bar{\theta}^{(j)})$, and*

$$\int_{\underline{\theta}^{(j)}}^{\bar{\theta}^{(j)}} Q_\alpha(\theta)\,\mathrm{d}F(\theta) = \int_{\underline{\theta}^{(j)}}^{\bar{\theta}^{(j)}} Q_{\mathrm{E}}(\theta)\,\mathrm{d}F(\theta);$$

3. efficient region*, where $Q_\alpha(\theta) = Q_{\mathrm{E}}(\theta) > U_\alpha(\theta)$ and $U'_\alpha(\theta) = \eta$ for any type $\theta \in (\underline{\theta}^{(j)}, \bar{\theta}^{(j)})$.*

Each of these three regions is a union of (potentially countably many) intervals. If some of the highest k types among the n agents, say $0 < \ell \le k$ of them, fall into one of the no-tension intervals, then each of the highest ℓ types receives one item and exert no effort. Similarly, if some of the highest k types among the n agents, say $0 < \ell \le k$ of them, fall into one of the efficient intervals, then each of the highest ℓ types receives one item and exert positive effort. However, if some of the highest $\ell > k$ types among the n agents fall into the same the no-effort interval, then no agent in this region exerts effort and the items are allocated randomly among the ℓ agents, meaning the highest k types receive one item each with probability 1, but the higher type has higher probability

[21] Existence of the optimal allocation rule is guaranteed by the compactness of the constraint set and continuity of the objective functional.

[22] If the partition is finite, say, the partition only consists of K disjoint intervals, define $\underline{\theta}^{(j)} = \bar{\theta}^{(j)}, \forall j > K$.

[23] The definitions of the interim allocation and utility on the cutoff points $\{\underline{\theta}^{(j)}\}_{j=1}^{\infty}$ do not affect the objective value.

Table 1: Three categories of regions under optimal contest

	(IC) binds	(IF) binds	effort	Q_α
no-tension region	$\times$	✓	$=0$	$=Q_{\mathrm{E}}$
no-effort region	✓	$\times$	$=0$	$\neq Q_{\mathrm{E}}$
efficient region	✓	✓	>0	$=Q_{\mathrm{E}}$

of receiving one item. In Section 7, we provide an example where we characterize the optimal interim allocation rule (see Fig. 3) and illustrate the allocation of the item (see Fig. 4).

The proof is provided in Appendix A.2 using tools from optimal control. Intuitively, we can view the principal's problem as an optimization problem with two inequality constraints ($\widehat{\mathrm{IF}}$) and (IC). Under optimality, either one of the constraints binds or neither of them binds. When ($\widehat{\mathrm{IF}}$) binds, optimal allocation rule and efficient allocation rule coincide. If the slope of the efficient allocation rule is no larger than the marginal cost, the efficient allocation rule can be implemented when no agent exerts effort. This happens in the no-tension region. However, if the slope of the efficient allocation rule is no larger than the marginal cost, (IC) requires that agents have to exert positive effort. This happens in the efficient region. In the second case, the principal can also consider the option of not allocating the items efficiently, e.g., ($\widehat{\mathrm{IF}}$) is slack, so that agents do not have incentives to exert effort, e.g., (IC) binds, which would imply that optimal allocation and utility coincide. This happens in the no-effort region. This is summarized in Table 1.

In general, both the number of each type of the intervals, and the order of each type of the intervals depend on the shape of the efficient allocation rule, and the coefficient α. In the following sections, we derive a sharper characterization on the number and the order of the intervals when the number of agents is sufficiently large.

6 Indirect implementation: Coarse Ranking Contests

In previous sections, we view a contest rule as a mapping from the signal profile to the allocation profile. However, in some papers in the literature of contest design (e.g., Moldovanu and Sela, 2001; Lazear and Rosen, 1981; Skaperdas, 1996), there is a subtler requirement on this mapping: a contest allocates the resource or the prize based *solely* on agents' rankings instead of the cardinal values of their signals.[24] In this section, we propose the concept of *coarse ranking*. In the coarse ranking, segment of signals are pooled and assigned the same coarse ranking. Correspondingly, we

[24]In the definition of Skaperdas (1996), Tullock contest allocates the resource stochastically based on the contest success function, which depends on the cardinal values of the whole signal (performance) profile. However, Konrad et al. (2009) pointed out that Tullock contest can be alternatively viewed as allocating the resource based on the ranking of a noisy signal s, described as $\log s = \log g(e) + \epsilon$, for some deterministic and strictly increasing function g and stochastic noise term ϵ, given effort e.

extend the usual interpretation of contest (or contest rule) to the so-called *coarse ranking contest* (or coarse ranking contest rule), where resources are allocated to the k agents with highest coarse ranking, with ties broken uniformly randomly. Our notion of coarse ranking contest generalizes the commonly used notion of contests in the literature.

We first provide a formal definition of coarse ranking.

Definition 5 (Coarse ranking)**.** *Given any countable set $\{(\underline{s}^{(j)}, \bar{s}^{(j)})\}_{j=1}^{\infty}$, the union of which is a subset of the type space, the coarse ranking of agent i under the performance profile $\boldsymbol{s} = (s_1, \ldots, s_n)$ is*

$$r_i(\boldsymbol{s}) = \left|\left\{i' \neq i, 1 \leq i' \leq n : s_{i'} > \bar{s}^{(j_{s_i})}\right\}\right|,$$

and the number of ties for agent i is

$$z_i(\boldsymbol{s}) = \left|\left\{i' \neq i, 1 \leq i' \leq n : \bar{s}^{(j_{s_{i'}})} = \bar{s}^{(j_{s_i})}\right\}\right| + 1,$$

where for any signal s_i, j_{s_i} is the interval j such that $s_i \in (\underline{s}^{(j)}, \bar{s}^{(j)})$ if s_i falls into one of the intervals within which signals are pooled to be assigned the same coarse ranking, and (by slightly overloading notation) denote $\bar{s}^{(j_{s_i})} = \underline{s}^{(j_{s_i})} = s_i$ if s_i is outside any of the intervals where signals are pooled. We call the functions $(\mathbf{r}, \mathbf{z})$ a coarse ranking.[25]

Moreover, when $\{(\underline{s}^{(j)}, \bar{s}^{(j)})\}_{j=1}^{\infty}$ is empty, the induced $(\mathbf{r}, \mathbf{z})$ defines a strict ranking.

Intuitively, each open interval in the given countable set of disjoint open intervals $\{(\underline{s}^{(j)}, \bar{s}^{(j)})\}_{j=1}^{\infty}$, specifies a region of signals that are pooled to be assigned the same (coarse) ranking. Outside the closure of any of this interval, signals are ranked strictly. When $\{(\underline{s}^{(j)}, \bar{s}^{(j)})\}_{j=1}^{\infty}$ is empty, the definition of coarse ranking coincides with the usual strict ranking where each agent i's ranking in a performance profile is just the order of agent i's signal in the given performance profile.

For any coarse ranking $(\mathbf{r}, \mathbf{z})$ and for any agent i, define the induced (coarse ranking) contest rule as

$$\tilde{x}_i(\boldsymbol{s}; \mathbf{r}, \mathbf{z}) = \begin{cases} 1 & k \geq r_i(\boldsymbol{s}) + z_i(\boldsymbol{s}) \\ \frac{k - r_i(\boldsymbol{s})}{z_i(\boldsymbol{s})} & k \in (r_i(\boldsymbol{s}), r_i(\boldsymbol{s}) + z_i(\boldsymbol{s})) \\ 0 & k \leq r_i(\boldsymbol{s}), \end{cases}$$

Definition 6 (Coarse ranking contest (rule))**.** *Any mapping profiles $\boldsymbol{x} : S \to X$ is a (coarse ranking) contest rule if there exists a coarse ranking $(\mathbf{r}, \mathbf{z})$ such that $x_i(\boldsymbol{s}) = \tilde{x}_i(\boldsymbol{s}; \mathbf{r}, \mathbf{z})$ for each i.*

By generalizing the concept of ranking, we define a larger class of (coarse ranking) contest rules, which is a subset of the class of mappings from performance space to allocation space. This set

[25] Intuitively, given signal profile $\boldsymbol{s}$, $r_i(\boldsymbol{s})$ is the number of the agents with rankings strictly above agent i, and $z_i(\boldsymbol{s})$ is the number of the agents with rankings tied agent i.

also includes the class of contest rules that allocate prize to agents based on the strict ranking of each agent in any given performance profile, examples of which include the three commonly studied contests in the literature: (1) all-pay contest, where the prize is allocated based on the ranking of agents' effort; (2) Lazear-Rosen contest, where the prize is allocated based on the ranking of a noisy signal of agents' effort, i.e., $s = e + \epsilon$ where e is effort, and ϵ is the stochastic noise; and (3) Tullock contest, where the prize is allocated based on the ranking of a noisy signal s of agents' effort, where the noisy signal is described by $\log s = \log g(e) + \epsilon$, with g being some strictly increasing function and ϵ being the stochastic noise (Fu and Wu, 2019; Konrad et al., 2009).

The next proposition provides a consistency check for Definition 2: using results in Kleiner et al. (2021), we show that when the interim allocation rule is symmetric, the mapping from signal space to the allocation space in this definition is, in a slightly broader sense, indeed a "contest" as termed in the economics literature.

Proposition 3. *Any symmetric interim allocation-utility pair (Q, U) that is implementable by a contest, has an indirect implementation that is a randomization over the coarse ranking contests.*

Proof. For any symmetric interim allocation-utility pair (Q, U) that is implementable by a contest, by definition, there exists a mapping from the signal space to the allocation space $x(\hat{s})$ specifying the allocation for each agent given the generated signal $\hat{s}$.[26] And the distribution over types $F(\theta)$ induces a distribution over signals, call it $\hat{F}(s)$. Similarly, a feasibility constraint on contest rule defined in the signal space could be induced from $(\widehat{\text{IF}})$. Such operations are valid since the recommended signal is a non-decreasing function of type. Following Theorem 1 and 2 in Kleiner et al. (2021), any monotone feasible contest rule x can be written as a convex combination of the extreme points. Using the construction of the extreme points in Theorem 3 in Kleiner et al. (2021), one can easily verify that the extreme points are the coarse ranking contest rules defined above. Hence x can be expressed as a randomization over the coarse ranking contest rules. Notice that agents' incentives are not affected during the above operations, and hence the expected utility of each agent in the coarse ranking contest is still U. □

7 Large Contests

In many applications of interests, the number of agents participating in the contest is large. In this section, we study large contests, and show that optimal contests exhibit simple structures in these settings. To simplify the exposition, we make the following assumption on top of Assumption 1 throughout the section. The main economic insights extend without this assumption.

[26] Notice that such a mapping might not exist if (Q, U) is implementable by a general mechanism that is not a contest.

Assumption 2 (Continuity)**.** *There exists* $\underline{\beta}_1, \bar{\beta}_1, \beta_2 \in (0, \infty)$ *such that* $f(\theta) \in [\underline{\beta}_1, \bar{\beta}_1]$ *and* $f'(\theta) \geq -\beta_2$ *and any type* $\theta \in [\underline{\theta}, \bar{\theta}]$*.*

7.1 Scarce Resources

In applications such as awarding presidential fellowships to students in universities, the competition is fierce and the ratio between the number of competing agents and the amount of awards is large. In this subsection, we consider the special case where $k = 1$, and the number of agents n is sufficiently large.[27] The efficient allocation rule becomes convex when the number of agents is sufficiently large, which simplifies the characterization of the optimal contest.

Lemma 5. *Let* $k = 1$*. Under Assumption 1 and 2, there exists a sufficiently large number* N *such that for any* $n \geq N$*, the efficient allocation rule* $Q_{\mathrm{E}}(\theta)$ *is convex in* θ*.*

Proof. Taking the second order derivative gives us

$$\begin{aligned} Q''_{\mathrm{E}} &= (F^{n-1})'' = ((n-1)F^{n-2} \cdot f)' = (n-1)((n-2)F^{n-3} \cdot f^2 + F^{n-2} \cdot f') \\ &\geq (n-1)F^{n-3}((n-2)\underline{\beta}_1^2 - F \cdot \beta_2) \geq 0 \end{aligned}$$

when $n \geq N \geq 2 + \frac{\beta_2}{\underline{\beta}_1^2}$. □

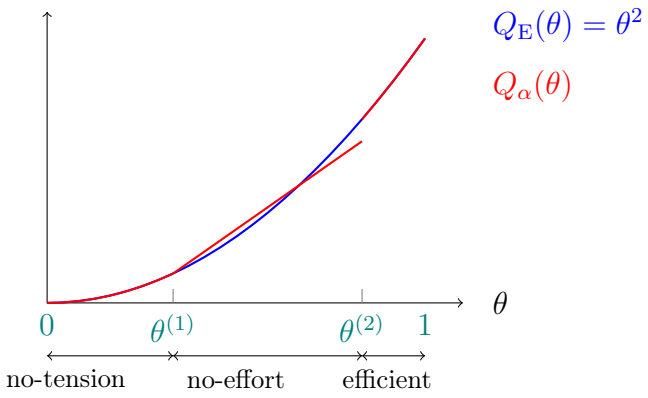
Convex Efficient Allocation Consider the case when the efficient allocation rule is convex. The optimal contest is characterized below.

Proposition 4. *Suppose* $Q_{\mathrm{E}}(\theta)$ *is convex in* θ*. Under Assumption 1, for any* $\alpha \in (0, 1)$*, there exists cutoff types* $\underline{\theta} \leq \theta^{(1)} \leq \theta^{(2)} \leq \bar{\theta}$ *such that the optimal contest* Q_α *divides the type space of each agent into at most three intervals, where* $(\underline{\theta}, \theta^{(1)})$ *is the no-tension region,* $(\theta^{(1)}, \theta^{(2)})$ *is the no-effort region, and* $(\theta^{(2)}, \bar{\theta})$ *is the efficient region.*

An example of the optimal interim allocation rule when the efficient allocation is convex is provided in Figure 3 in Section 2. The contest rule that implements this optimal allocation rule is provided in Figure 4 in Section 2. Intuitively, when the efficient allocation is convex, the derivative of the efficient allocation cross η from below only once. Therefore, for small types, there is no tension since the derivative of the efficient allocation is sufficiently small, and the optimal contest can allocate the item efficiently without requiring effort from the agents. For high types, since the change in efficient allocation is sufficiently large, the incentive constraints bind and the interim utility must be a linear function. Moreover, in order for the interim allocation to be interim feasible,

[27] The analysis for k being a small constant and n being sufficiently large is significantly more involved. We omit it in our paper since the economic insights are similar.

Figure 3: Optimal interim allocation rule under convex $Q_{\mathrm{E}}(\theta)$



†Suppose $n = 2$, $k = 1$, $F(\theta) = \theta^2$, $\theta \in [0, 1]$ and $\eta = 1$.

†The interim efficient allocation rule is $Q_{\mathrm{E}}(\theta) = F^{n-1}(\theta) = \theta^2$ in this example, i.e., the highest type gets the item. $Q_\alpha(\theta)$ is the optimal interim allocation rule.

in this linear utility region, it must be the case that no-effort region occurs before the efficient region, not the other way around. The formal proof of Proposition 4 is provided in Appendix A.3.

Interestingly, in optimal contests, when the efficient allocation rule is convex, there is no distortion at both the bottom and the top, while there are distortions for middle types, which leads to a sharp contrast to the classical auction design setting where there are distortions at the bottom.

Convergence Results By applying the characterization of optimal contests for convex efficient allocation rules and Lemma 5, we immediately obtain the characterization of optimal contests for the allocation of scarce resource. Moreover, we show that when the number of agents increases, the no-tension region converges to the full type space. Since the contest format for the no-tension region is winner-takes-all, this further implies that in the limiting case, the format of the optimal contest is essentially winner-takes-all.

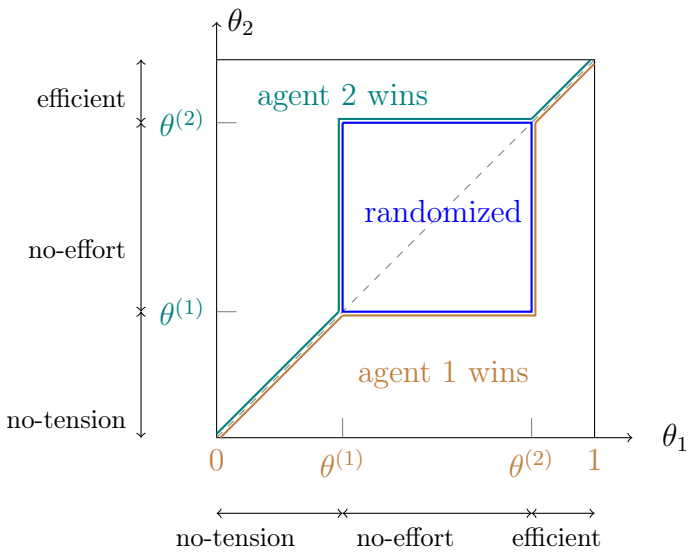
Theorem 3 (Convergence of Contest Format)**.** *Let $k = 1$. Under Assumption 1 and 2, for any $\alpha \in (0, 1)$, there exists N such that for any $n \geq N$, the optimal contest takes the form described in Proposition 4. Moreover, the no-tension region converges to the entire type space as n goes to infinity.*

Proof. By Lemma 5, for sufficiently large n, the efficient allocation rule is convex. Therefore, the interim allocation rule of the optimal contests takes the form of Proposition 4.

Let $Q_{\alpha,n}(\theta)$ and $Q_{\mathrm{E},n}(\theta)$ be the optimal interim allocation rule and efficient allocation rule in a contest with $n < \infty$ agents. For any finite n, we have that

$$\frac{1}{n} \geq \int_{\theta_n^{(1)}}^{\bar{\theta}} Q_{\mathrm{E},n}(\theta)\,\mathrm{d}F(\theta) \geq \int_{\theta_n^{(1)}}^{\bar{\theta}} \left(\eta \cdot (\theta - \theta_n^{(1)}) + Q_{\mathrm{E},n}(\theta_n^{(1)})\right)\,\mathrm{d}F(\theta).$$

Figure 4: Implementation of the optimal allocation rule



†When both agents produce signals in $(\theta^{(1)}, \theta^{(2)})$, the item is allocated randomly, but the agent with the higher signal has a higher probability that is strictly less than 1, of getting the item. When there is at least one agent who produces a signal outside $(\theta^{(1)}, \theta^{(2)})$, the item is allocated to the agent with a higher signal.

†Suppose $n = 2$, $k = 1$, $F(\theta) = \theta^2$, $\theta \in [0, 1]$.

The first inequality holds since the ex ante probability each agent gets the item is at most $\frac{1}{n}$, and the second inequality holds since the efficient allocation majorizes the interim allocation, where the latter is again at least the interim utility. Since $Q_{\mathrm{E},n}(\theta_n^{(1)})$ is non-negative, we have that

$$\int_{\theta_n^{(1)}}^{\bar{\theta}} (\theta - \theta_n^{(1)}) \, \mathrm{d}F(\theta) \leq \frac{1}{n\eta}$$

for any n. Note that $\frac{1}{n\eta}$ converges to 0 when n converges to infinity. In order for the inequality to hold, $\theta_n^{(1)}$ must also converge to $\bar{\theta}$ as n converges to infinity. □

Given the simplicity of the winner-takes-all contest and the convergence result, in practice, it is tempting to use the winner-takes-all contest as an approximate to the optimal contest, even when the number of agents is finite. However, we will show that principal's payoff under the winner-takes-all contest does not converge to her payoff under optimal contest as the number of agents increases. This is because under the optimal contest, by randomizing the allocation for a small range of high types, the principal could significantly improve the agents' expected utilities while keeping the loss in matching efficiency small. For any interim allocation rule Q that is implementable by a contest, by Definition 2 and Proposition 1, there exists a unique interim utility U with $U(\underline{\theta}) = Q(\underline{\theta})$ such that (Q, U) is implementable by a contest and achieves weakly higher objective value for the principal

than any other mechanism (Theorem 1). Denote this value by $V_\alpha(Q)$, i.e.,

$$V_\alpha(Q) = \sup\{\mathrm{Obj}_\alpha(Q,U) : U(\underline{\theta}) = Q(\underline{\theta}) \text{ and } (Q,U) \text{ is implementable by a contest}\}$$

Theorem 4 (Non-convergence in Payoffs)**.** *Let $k = 1$. Under Assumption 1 and 2, for any $\alpha \in (0,1)$ and any sufficiently small $\epsilon > 0$, there exists $N_{F,\epsilon}$ such that for any finite $n > N_{F,\epsilon}$, the ratio between the objective value of the optimal contest and the winner-takes-all contest is at least $\delta \triangleq \frac{(\bar{\theta}-\epsilon)\cdot\alpha+1-\alpha}{\bar{\theta}\cdot\alpha+(1-\alpha)(1-\frac{1}{e}+\epsilon)} > 1$, i.e., $\frac{V_\alpha(Q_{\alpha,n})}{V_\alpha(Q_{\mathrm{E},n})} \geq \delta$.*

Proof of Theorem 4. First we present Lemma 6, whose proof can be found in Appendix A.3. It shows that given the efficient allocation rule, the sum of expected utility of each agent is small compared to the first best of 1, i.e., the highest type getting the item without exerting effort.

Lemma 6. *For any $\epsilon > 0$, there exists $N_0 \geq 1$ such that for any $n \geq N_0$, we have $n \cdot \mathbf{E}_{\theta\sim F}[U_{\mathrm{E},n}(\theta)] \leq 1 - \frac{1}{e} + \epsilon$.*

Intuitively, the competition among agents is high for agents with sufficiently high types. Thus agents with high types need to exert high efforts to ensure a large allocation, leading to a utility loss compared to the first best utility. By applying Lemma 6, we obtain an upper bound on the performance of the winner-takes-all contest. That is, for any $\epsilon > 0$, there exist N_0 such that for any $n \geq N_0$, we have

$$\begin{aligned} n \cdot V_\alpha(Q_{\mathrm{E},n}) &= n\alpha \cdot \mathbf{E}_{\theta\sim F}[\theta \cdot Q_{\mathrm{E},n}(\theta)] + n(1-\alpha) \cdot \mathbf{E}_{\theta\sim F}[U_{\mathrm{E},n}(\theta)] \\ &\leq \alpha \cdot \bar{\theta} + (1-\alpha) \cdot \left(1 - \frac{1}{e} + \epsilon\right). \end{aligned}$$

The inequality holds due to Lemma 6 and the fact the the upper bound of type of the agent winning the item is $\bar{\theta}$.

Next we provide a lower bound on the performance of the optimal contest. In particular, for any n large enough, consider a feasible allocation

$$Q_n(\theta) = \begin{cases} Q_{\mathrm{E},n}(\theta) & \text{if } \theta \leq \hat{\theta}_n; \\ \eta \cdot (\theta - \hat{\theta}_n) + Q_{\mathrm{E},n}(\hat{\theta}_n) & \text{if } \theta > \hat{\theta}_n, \end{cases}$$

such that $\mathbf{E}_{\theta\sim F}[Q_n(\theta)] = \mathbf{E}_{\theta\sim F}[Q_{\mathrm{E},n}(\theta)] = \frac{1}{n}$. Let $U_n(\theta) = Q_n(\theta)$. Notice that such that (Q_n, U_n) satisfies the IC constraints. Moreover, $Q_n(\theta)$ induces no effort and hence $\mathbf{E}_{\theta\sim F}[U_n(\theta)] = \frac{1}{n}$. In the following lemma, we show that the matching efficiency of the given contest rule converges to the optimal welfare when the number of agents is sufficiently large. The proof of Lemma 7 is provided in Appendix A.3.

Lemma 7. *For any $\epsilon > 0$, there exists N_1 such that for any $n \geq N_1$, $n \cdot \mathbf{E}_{\theta\sim F}[\theta \cdot Q_n(\theta)] \geq \bar{\theta} - \epsilon$.*

Therefore, there exists N_1 such that for any $n \geq N_1$, we have

$$\begin{aligned} n \cdot V_\alpha(Q_{\alpha,n}) &\geq n \cdot \alpha \mathbf{E}_{\theta\sim F}[\theta \cdot Q_n(\theta)] + n \cdot (1-\alpha)\mathbf{E}_{\theta\sim F}[U_n(\theta)] \\ &\geq \alpha(\bar{\theta} - \epsilon) + 1 - \alpha. \end{aligned}$$

Finally, for any $\epsilon > 0$, letting $N = \max\{N_0, N_1\}$, for any $n \geq N$, by combining the inequalities above, we have

$$\frac{V_\alpha(Q_{\alpha,n})}{V_\alpha(Q_{\mathrm{E},n})} \geq \frac{(\bar{\theta} - \epsilon) \cdot \alpha + 1 - \alpha}{\bar{\theta} \cdot \alpha + (1-\alpha)(1 - \frac{1}{e} + \epsilon)}.$$

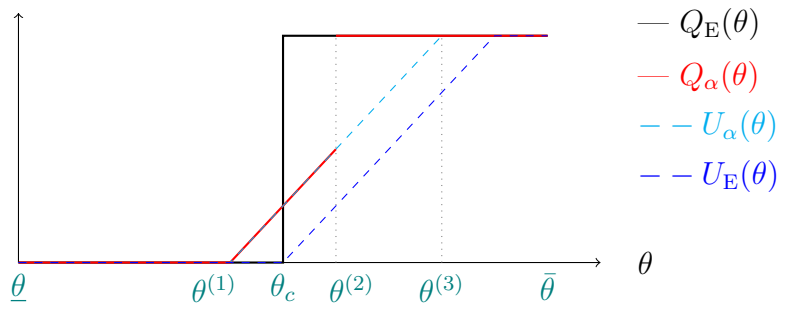
□

Note that our framework enables us to completely characterize the optimal contest in large but finite-size contests, as opposed to Olszewski and Siegel (2016, 2020), which uses a continuum model to approximate equilibrium in contests. A striking observation derived through the characterization for arbitrary finite-size contests is that although the contest format converges as the number of agents increases, which is consistent with Olszewski and Siegel (2016), the optimal payoff may not converge to the payoff given the contest format in the limit. Specifically, in our application of allocation with scarce resource, the limit contest format is winner-takes-all, but the principal can significantly improve the objective value by randomizing the allocation within a small region of high signal realizations.

This non-convergence result contrasts with Bulow and Klemperer (1996), who shows that by running the second price auction with $n+1$ agents, which is simple and implements the efficient allocation rule, the principal could obtain higher objective value compared to running the optimal mechanism with n agents. In our setting, winner-takes-all contest is the simple mechanism that maximizes matching efficiency. However, the loss in objective value coming from this simple mechanism can never be compensated, no matter how many additional agents are added.[28] The economic intuition comes from the ambiguous effect of competition in our setting. In Bulow and Klemperer (1996), allocating the items efficiently does not necessarily oppose the principal's objective, while in our setting, the principal's objective function contains two opposing interests: to achieve matching efficiency, higher competition is better but high competition strongly incentivizes effort, which lowers agents' utilities. The WTA contest maximizes competition among agents, while the optimal contest shall be the one with a competition level balancing the two different interests in principal's objective. This explains why the utility loss in WTA contest cannot be recovered by the increase in matching efficiency no matter how many additional agents are added.

[28] A caveat of this comparison is that in Bulow and Klemperer (1996), the principal maximizes revenue, while in our paper, the principal maximizes welfare, which is defined as the weighted average between matching efficiency and the sum of agents' utilities.

Figure 5: Optimal allocation and utility for large scale economy in the limit case.



Any type in $(\theta^{(1)}, \theta^{(2)})$ is called middle types, whose utilities are improved under the optimal contest because they do not exert effort. Types above $\theta^{(2)}$ are called high types, whose utilities are weakly improved and strictly for some, because the effort each has to exert is lower.

7.2 Large Scale Economy

In applications such as college admission and affordable housing, the amount of resources is not necessarily scarce. To capture such applications, consider a setting with n agents and $0 < k < n$ items, and replicate both the agents and the items $z \in \mathbb{N}_+$ times. The parameter z captures the scale of the economy. When the scale z goes to infinity, the efficient allocation rule in this setting converges to the cutoff rule, where the top $\frac{k}{n}$ fraction of the types are allocated with the resource. Allocating the items efficiently hence creates strong incentives for types close to the cutoff to exert costly effort. Theorem 5 shows that in the optimal contest, the items are allocated to types close to the cutoff randomly so that there is no incentive for those types to exert effort. Let θ_c be the cutoff type such that $\Pr[\theta \geq \theta_c] = \frac{k}{n}$.

Theorem 5. *Under Assumption 1 and 2, for any $\alpha \in (0, 1)$, and any fixed integers $n > k > 0$, there exists Z such that for any integer $z \geq Z$, in a setting with $z \cdot n$ agents and $z \cdot k$ items, there exists cutoff types $\underline{\theta} \leq \theta^{(1)} < \theta_c < \theta^{(2)} \leq \theta^{(3)} \leq \bar{\theta}$ so that the optimal contest divides the type space into at most four intervals, where $(\underline{\theta}, \theta^{(1)})$ is the no-tension region; $(\theta^{(1)}, \theta^{(2)})$ is the no-effort region; $(\theta^{(2)}, \theta^{(3)})$ is the efficient region; and $(\theta^{(3)}, \bar{\theta})$ is the no-tension region.*

Intuitively, in the limit, the efficient allocation converges to a step function, where only types above θ_c receive the items. The interim utility given the efficient allocation rule is illustrated in Figure 5 as the blue curve. In order to improve the weighted average between the matching efficiency and the sum of expected utilities, the principal can randomize the allocation around the cutoff type θ_c, i.e., randomizing the allocation for types within $(\theta^{(1)}, \theta^{(2)})$. This lead to an efficiency loss of at most $\theta^{(2)} - \theta^{(1)}$ when an item is allocated inefficiently. However, given this randomization in allocation, the utilities for a wider range of types are improved compared to the efficient allocation rule, i.e., utilities for types within $(\theta^{(1)}, \theta^{(3)})$ are improved. The improvement in expected utility

can be significantly larger than the efficiency loss when $\theta^{(1)}$ is sufficiently close to $\theta^{(2)}$. Therefore, such no-effort region always exists in the optimal contest to improve the principal's objective value. Finally, for types that are sufficiently low or sufficiently high, it is easy to verify that both the matching efficiency and the sum of expected utilities are maximized when the items are allocated efficiently. In Appendix A.4, we show that this intuition extends when the scale of the economy is finite but sufficiently large.

Our result is reminiscent of the Director's law, i.e., public programs should be designed primarily for the benefit of middle classes. Although principal cares about the utilities of all agents, the optimal contest gives "preferential treatments" to the middle types, in the sense that the middle types are given higher utilities in our setting compared to in a fully competitive setting where items are allocated efficiently. This is optimal to the principal because by giving "preferential treatments" to the middle types, they exert no effort, which weakly (strictly) decreases the effort level for all (some) types above. The empirical result in Krishna et al. (2022) is in line with this logic. They use data in Turkey to show that randomly allocating the college seats to students, especially those with low scores, reduces the stress for all students. They also theoretically analyze when this is optimal. The difference is our setting is that instead of randomizing allocation to low types, we predict that it is optimal to randomize allocation to types around the cutoff.

8 Discussions

In this section, we provide several extensions of our model and discuss the further connection of our results to the literature.

8.1 Implementation of Non-monotone Allocation Rules

In this paper, we have shown that it is without loss of optimality to focus on contests with monotone interim allocations. Here we show that there exists non-monotone interim allocations that can also be implemented as contests. The main intuition is that the expected allocation given a wide range of signal realizations could be indifferent for the agent with different types, and hence it is possible for a high type to choose a low signal realization with low expected allocation.

Example 3. *Consider a simplified single-agent setting where the agent's type θ is drawn from uniform distribution in $[0,1]$, and the ability of the agent is $\eta = 1$. Suppose the contest rule is $x(s) = \min\{1, s\}$. For any type $\theta \in [0,1]$, the agent is indifferent generating any signals between $[\theta, 1]$. One feasible equilibrium strategy for the agent is to choose signal $s(\theta) = 1 - \theta$ if $\theta \leq \frac{1}{2}$, and $s(\theta) = \theta$ otherwise. It is easy to verify that the induced interim allocation rule is $Q(\theta) = 1 - \theta$ if $\theta \leq \frac{1}{2}$, and $Q(\theta) = \theta$ otherwise. This interim allocation rule is strictly decreasing for type in $[0, \frac{1}{2}]$, and strictly increasing for type in $[\frac{1}{2}, 1]$.*

8.2 Concave Pareto Frontier and Linear Objective

In this paper, we have focused on the objective of maximizing the weighted average between matching efficiency and agents' expected utilities. An alternative is to consider the Pareto optimal contest for these two objectives. We show that the Pareto frontier of this problem is concave. The concavity of the Pareto frontier justifies that it is without loss of generality to assume that the principal's objective function is a convex combination of matching efficiency and agents' expected utilities. Moreover, it shows that there is no gain in randomizing over allocation rules.

Given any constant $U \geq 0$ as the lower bound of the agents' expected utilities, we define the Pareto frontier $E(U)$ as

$$\begin{aligned} E(U) = \quad & \sup_{\boldsymbol{Q},\boldsymbol{U}} \mathbf{E}_{\boldsymbol{\theta}}\left[\sum_{i=1}^{n} \theta_i \cdot Q_i(\theta_i)\right] \\ \text{s.t.} \quad & (\boldsymbol{Q},\boldsymbol{U}) \text{ is implementable by a contest,} \\ & \mathbf{E}_{\boldsymbol{\theta}}\left[\sum_{i=1}^{n} U_i(\theta_i)\right] \geq U. \end{aligned} \tag{PF}$$

Proposition 5. *The Pareto frontier $E(U)$ is concave.*

Proof. Consider any $U', U'' \geq 0$, $\lambda \in [0,1]$. Suppose $(\boldsymbol{Q}', \boldsymbol{U}')$ and $(\boldsymbol{Q}'', \boldsymbol{U}'')$ attain the value $E(U')$ and $E(U'')$ defined in program (PF).

Note that for any $\lambda \in [0,1]$, $\lambda\boldsymbol{Q}' + (1-\lambda)\boldsymbol{Q}''$ is still interim feasible. Although the allocation-utility pair $(\lambda\boldsymbol{Q}' + (1-\lambda)\boldsymbol{Q}'', \lambda\boldsymbol{U}' + (1-\lambda)\boldsymbol{U}'')$ may violate the IC constraints, by 1, there exists another utility profile $U^\dagger$ such that (1) $U_i^\dagger(\theta_i) \geq \lambda U_i'(\theta_i) + (1-\lambda)U_i''(\theta_i)$ for any agent i and any type θ_i; and (2) $(\lambda\boldsymbol{Q}' + (1-\lambda)\boldsymbol{Q}'', U^\dagger)$ is implementable by a contest. This implies that $\lambda\boldsymbol{Q}' + (1-\lambda)\boldsymbol{Q}''$ is a feasible solution to program (PF) when $U = \lambda U' + (1-\lambda)U''$. Therefore, $E(\lambda U' + (1-\lambda)U'') \geq \lambda E(U') + (1-\lambda)E(U'')$. □

8.3 Cost for Downward Deviation

In this paper, we have assumed that the agent bears no cost for generating signals lower than his true type. This assumption has no bite in our model, and all of the results directly generalize when introducing positive costs for downward deviations. The main reason is that agents will not have incentives to generate signals lower than their true types in optimal contests in order to maximizes their expected utilities. By adding positive costs for downward deviations, the incentive constraints for downward deviations remain slack, and the optimal contests remain unchanged.

8.4 Asymmetric Environments

Most of the results in our paper are stated for symmetric environments to simplify the exposition. The main insights derived through our analysis (e.g., Theorem 2–5) extend to asymmetric environments where agents are heterogeneous ex ante. Intuitively, even in asymmetric environments, the Boarder's feasibility constraint can be decomposed into n separate majorization constraints on agents' interim allocations. Combined with the IC constraints for each individual agent, in the optimal contest for asymmetric environments, the optimal interim allocation rule still features three different types of regions, with the partitions of the type space being different across agents.

A slight difference in asymmetric environments is that the contest format for implementing a given feasible interim allocation rule may be different. For example, the efficient allocation rule can also be implemented in asymmetric environments. However, in contrast to symmetric environments, in the case of $k = 1$, the contest rule may not take the form of winner-takes-all.

Example 4. *Consider the setting with $n = 2$, $\Theta_1 = \Theta_2 = [0, 1]$, $\eta_1 = \eta_2 = 1$ and concave CDFs $F_1 \underset{FOSD}{\succsim} F_2$. The efficient allocation is implementable by the following contest: $x_i(s_i, s_j) = 1$ if $F_i^{-1}(s_i) > F_j^{-1}(s_j)$, and is zero otherwise.*[29] *Note that this is not a winner-takes-all contest.*

References

Askenazy, P., Breda, T., Moreau, F., and Pecheu, V. (March 2022). Do french companies underreport their workforce at 49 employees to get around the law? Technical Report policy brief no. 82, Institute des politiques publiques.

Ball, I. (2019). Scoring strategic agents. *arXiv preprint arXiv:1909.01888*.

Barut, Y. and Kovenock, D. (1998). The symmetric multiple prize all-pay auction with complete information. *European Journal of Political Economy*, 14(4):627–644.

Baye, M. R., Kovenock, D., and De Vries, C. G. (1993). Rigging the lobbying process: an application of the all-pay auction. *The American Economic Review*, 83(1):289–294.

Ben-Porath, E., Dekel, E., and Lipman, B. (2023). Sequential mechanisms for evidence acquisition. *working paper*.

Ben-Porath, E., Dekel, E., and Lipman, B. L. (2014). Optimal allocation with costly verification. *American Economic Review*, 104(12):3779–3813.

[29] In this contest, each agent i' equilibrium signal choice is $\hat{s}_i(\theta_i) = F_i(\theta_i)$.

Border, K. C. (1991). Implementation of reduced form auctions: A geometric approach. *Econometrica: Journal of the Econometric Society*, pages 1175–1187.

Bulow, J. and Klemperer, P. (1996). Auctions versus negotiations. *The American Economic Review*, 86(1):180–194.

Chawla, S., Hartline, J. D., and Sivan, B. (2019). Optimal crowdsourcing contests. *Games and Economic Behavior*, 113:80–96.

Che, Y.-K. and Gale, I. L. (1998). Caps on political lobbying. *The american economic review*, 88(3):643–651.

Che, Y.-K., Kim, J., and Mierendorff, K. (2013). Generalized reduced-form auctions: A network-flow approach. *Econometrica*, 81(6):2487–2520.

Clark, D. J. and Riis, C. (1998). Competition over more than one prize. *The American Economic Review*, 88(1):276–289.

Clarke, F. (2013). *Functional analysis, calculus of variations and optimal control*, volume 264. Springer.

Conix, S., De Block, A., and Vaesen, K. (2021). Grant writing and grant peer review as questionable research practices. *F1000Research*, 10.

Fang, D., Noe, T., and Strack, P. (2020). Turning up the heat: The discouraging effect of competition in contests. *Journal of Political Economy*, 128(5):1940–1975.

Frankel, A. and Kartik, N. (2019). Muddled information. *Journal of Political Economy*, 127(4):1739–1776.

Fu, Q. and Wu, Z. (2019). Contests: Theory and topics. In *Oxford Research Encyclopedia of Economics and Finance*.

Gershkov, A., Moldovanu, B., Strack, P., and Zhang, M. (2022). Optimal insurance: Dual utility, random losses and adverse selection. *Revise and Resubmit at American Economic Review*.

Green, J. R. and Laffont, J.-J. (1986). Partially verifiable information and mechanism design. *The Review of Economic Studies*, 53(3):447–456.

Greenwald, A., Oyakawa, T., and Syrgkanis, V. (2018). Simple vs optimal contests with convex costs. In *Proceedings of the 2018 World Wide Web Conference*, pages 1429–1438.

Hardt, M., Megiddo, N., Papadimitriou, C., and Wootters, M. (2016). Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, pages 111–122.

Hartline, J. D. and Roughgarden, T. (2008). Optimal mechanism design and money burning. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 75–84.

Kartik, N. (2009). Strategic communication with lying costs. *The Review of Economic Studies*, 76(4):1359–1395.

Kleiner, A., Moldovanu, B., and Strack, P. (2021). Extreme points and majorization: Economic applications. *Econometrica*, 89(4):1557–1593.

Konrad, K. A. et al. (2009). Strategy and dynamics in contests. *OUP Catalogue*.

Krishna, K., Lychagin, S., Olszewski, W., Siegel, R., and Tergiman, C. (2022). Pareto improvements in the contest for college admissions. Technical report, National Bureau of Economic Research.

Lazear, E. P. and Rosen, S. (1981). Rank-order tournaments as optimum labor contracts. *Journal of political Economy*, 89(5):841–864.

Moldovanu, B. and Sela, A. (2001). The optimal allocation of prizes in contests. *American Economic Review*, 91(3):542–558.

Moldovanu, B. and Sela, A. (2006). Contest architecture. *Journal of Economic Theory*, 126(1):70–96.

Myerson, R. B. (1981). Optimal auction design. *Mathematics of operations research*, 6(1):58–73.

Mylovanov, T. and Zapechelnyuk, A. (2017). Optimal allocation with ex post verification and limited penalties. *American Economic Review*, 107(9):2666–94.

Olszewski, W. and Siegel, R. (2016). Large contests. *Econometrica*, 84(2):835–854.

Olszewski, W. and Siegel, R. (2020). Performance-maximizing large contests. *Theoretical Economics*, 15(1):57–88.

Perez-Richet, E. and Skreta, V. (2022). Test design under falsification. *Econometrica*, 90(3):1109–1142.

Perez-Richet, E. and Skreta, V. (2023). Fraud-proof non-market allocation mechanisms. *working paper*.

Sansone, R. A. and Sansone, L. A. (2011). Faking attention deficit hyperactivity disorder. *Innovations in Clinical Neuroscience*, 8(8):10.

Siegel, R. (2009). All-pay contests. *Econometrica*, 77(1):71–92.

Siegel, R. (2010). Asymmetric contests with conditional investments. *American Economic Review*, 100(5):2230–60.

Siegel, R. (2014). Asymmetric contests with head starts and nonmonotonic costs. *American Economic Journal: Microeconomics*, 6(3):59–105.

Skaperdas, S. (1996). Contest success functions. *Economic theory*, 7(2):283–290.

Spence, M. (1973). Job market signaling. *Quarterly Journal of Economics*, 87(3):355–374.

Zhang, M. (2023). Optimal contests with incomplete information and convex effort costs. *accepted at Theoretical Economics*.

A Missing Proof

A.1 Proof omitted in Section 4

Proof of Lemma 1. We will prove the if and only if conditions separately for each direction.

Only if: If $(\boldsymbol{Q}, \boldsymbol{U})$ is implementable by a contest, by Definition 2, there exist signal recommendation policy $\hat{\boldsymbol{s}}$ and contest rule $\boldsymbol{x}$ that induce $(\boldsymbol{Q}, \boldsymbol{U})$. $\boldsymbol{Q}$ satisfies interim feasibility because it is induced by the ex-post allocation rule $q_i(\boldsymbol{\theta}) = x_i(\hat{s}_i(\theta_i), \hat{\boldsymbol{s}}_{-i}(\theta_{-i}))$ for all i and $\boldsymbol{\theta}$. Notice that for any signal recommendation policy $\hat{\boldsymbol{s}}$ and contest rule $\boldsymbol{x}$ that implements $(\boldsymbol{Q}, \boldsymbol{U})$, it is without loss to assume that any realization of the recommendation $\hat{s}_i(\theta_i)$ is weakly higher than θ_i, This is because weakly increasing any signal recommendation below θ_i will not incur any effort cost for type θ_i, but weakly decreases other types' incentives for deviation.

For any agent i and any pair of types $\theta_i < \theta_i'$, let s_i' be the largest signal realization given $\hat{s}_i(\theta_i')$. Thus we have $s_i' \geq \theta_i'$. Note that agent i with type θ_i' has utility $U_i(\theta_i')$ for choosing signal s_i' as the agent must be indifferent for all his signal realizations. Therefore, agent i's utility for reporting signal s_i' when his type is θ_i is

$$\begin{aligned}\mathbf{E}_{\theta_{-i}}\left[x_i(s_i', \hat{s}_{-i}(\theta_{-i}))\right] - \eta \cdot e(s_i', \theta_i) &= \mathbf{E}_{\theta_{-i}}\left[x_i(s_i', \hat{s}_{-i}(\theta_{-i}))\right] - \eta \cdot e(s_i', \theta_i') - \eta \cdot (\theta_i' - \theta_i) \\ &= U_i(\theta_i') - \eta \cdot (\theta_i' - \theta_i).\end{aligned}$$

Since the agent has weakly lower utility for deviating his choice of signals, we have

$$U_i(\theta_i) \geq U_i(\theta_i') - \eta \cdot (\theta_i' - \theta_i).$$

By rearranging the terms and taking the limit $\theta_i' \to \theta_i$, we have $U_i'(\theta) \leq \eta$. Similarly, s_i be the largest signal realization given $\hat{s}_i(\theta_i)$. We have

$$\begin{aligned}U_i(\theta_i') &\geq \mathbf{E}_{\theta_{-i}}[x_i(s_i, \hat{s}_{-i}(\theta_{-i}))] - \eta \cdot e(s_i, \theta_i') \\ &\geq \mathbf{E}_{\theta_{-i}}[x_i(s_i, \hat{s}_{-i}(\theta_{-i}))] - \eta \cdot e(s_i, \theta_i) = U_i(\theta_i).\end{aligned}$$

Again by rearranging the terms and taking the limit $\theta_i' \to \theta_i$, we have $U_i'(\theta) \geq 0$, and hence $U_i'(\theta_i) \in [0, \eta]$ for any type θ_i.

Finally, as the effort is non-negative, the interim allocation must be weakly larger than the interim utility. When the inequality is strict, the agent must be choosing signal realizations strictly higher than his type with probability. In this case, we have $s_i > \theta_i$. For any type $\theta_i' \in (\theta_i, s_i)$ we

have

$$\begin{aligned} U_i(\theta_i') &\geq \mathbf{E}_{\theta_{-i}}[x_i(s_i, \hat{s}_{-i}(\theta_{-i}))] - \eta \cdot e(s_i, \theta_i') \\ &= \mathbf{E}_{\theta_{-i}}[x_i(s_i, \hat{s}_{-i}(\theta_{-i}))] - \eta \cdot e(s_i, \theta_i) + \eta \cdot (\theta_i' - \theta_i) \\ &= U_i(\theta_i) + \eta \cdot (\theta_i' - \theta_i). \end{aligned}$$

By rearranging the terms and taking the limit $\theta_i' \to \theta_i$, we have $U_i'(\theta_i) \geq \eta$. Since we also know that $U_i'(\theta_i) \leq \eta$, both inequalities must be equalities and hence $U_i'(\theta_i) = \eta$.

If: Since $\boldsymbol{Q}$ is interim feasible, there exists an ex post allocation rule $\boldsymbol{q}$ that implements $\boldsymbol{Q}$. Consider the signal recommendation policy $\hat{\boldsymbol{s}}$ where $\hat{s}_i(\theta_i) = \theta_i + \frac{1}{\eta}(Q_i(\theta_i) - U_i(\theta_i))$ for any agent i with type θ_i. It is easy to verify that $\hat{s}_i(\theta_i)$ is monotone in θ_i since $Q_i'(\theta_i) \geq 0$ and $U_i'(\theta_i) \leq \eta$. Let $\theta_i(s_i)$ be the inverse of function $\hat{s}_i$.[30] Consider the contest rule $\boldsymbol{x}$ where $x_i(\boldsymbol{s}) = q_i(\boldsymbol{\theta}(\boldsymbol{s}))$ for all agent i. We show that $\hat{\boldsymbol{s}}$ and $\boldsymbol{x}$ implements $(\boldsymbol{Q}, \boldsymbol{U})$.

First by our construction, when all agents follow the recommendation, both the interim allocation and the interim utility coincide with $\boldsymbol{Q}$ and $\boldsymbol{U}$ respectively. Then it is sufficient to show that the agents have weak incentives to follow the recommendation. In particular, for any agent i with type θ_i, by deviating the report to type $\theta_i' > \theta_i$, the utility for deviation is

$$Q_i(\theta_i') - \eta \cdot e(\hat{s}_i(\theta_i'), \theta_i) = U_i(\theta_i') - \eta \cdot (\theta_i' - \theta_i) \leq U_i(\theta_i)$$

where the last inequality holds since the derivative of U is always at most η. We analyze the incentives for downward deviation in three cases. If the deviation type $\theta_i' < \theta_i$ satisfies $Q_i(\theta_i') = U_i(\theta_i')$, the utility for deviation is

$$Q_i(\theta_i') - \eta \cdot e(\hat{s}_i(\theta_i'), \theta_i) = U_i(\theta_i') \leq U_i(\theta_i).$$

If the deviation type $\theta_i' < \theta_i$ satisfies $Q_i(\theta_i') > U_i(\theta_i')$, let $\theta_i^\dagger > \theta_i'$ be the smallest type such that $Q_i(\theta_i^\dagger) = U_i(\theta_i^\dagger)$. If $\theta_i \geq \theta_i^\dagger$, the utility for deviation is

$$Q_i(\theta_i') - \eta \cdot e(\hat{s}_i(\theta_i'), \theta_i) \leq Q_i(\theta_i^\dagger) = U_i(\theta_i^\dagger) \leq U_i(\theta_i).$$

If $\theta_i < \theta_i^\dagger$, the derivative of U for any type between θ_i' and θ_i must be constant η. Moreover, and hence the utility for deviation is

$$Q_i(\theta_i') - \eta \cdot e(\hat{s}_i(\theta_i'), \theta_i) \leq U_i(\theta_i') + \eta \cdot (\theta_i - \theta_i') = U_i(\theta_i).$$

[30] Note that $\hat{s}_i(\theta_i)$ is only weakly monotone. When there are multiple types θ_i with the same signal recommendation s_i, we map s_i randomly to those types according to the type distribution F_i.

By combining the inequalities, all agents have no incentives to deviate based on the recommendations. □

Proof of Lemma 2. We can view a general mechanism as offering a menu of randomized signal recommendations based on the reported type profiles. The partial derivative of agent i's utility with respect to his own type is between 0 and η. By envelope theorem, the derivative of the interim utility is between 0 and η. Finally, in any general mechanism we have $U_i(\theta_i) \leq Q_i(\theta_i)$ for all θ_i and all i because the effort cost of each agent is non-negative. □

Proof of Proposition 1. For any agent i, given any monotone and interim feasible allocation $\boldsymbol{Q}$, and $\underline{u}_i \leq Q_i(\underline{\theta}_i)$, let

$$U_i(\theta_i) = \min\left\{\underline{u}_i + \eta(\theta_i - \underline{\theta}_i),\ \inf_{\theta_i' \leq \theta_i} Q_i(\theta_i') + \eta(\theta_i - \theta_i')\right\}. \tag{4}$$

Notice that $U_i(\theta_i) \leq Q_i(\theta_i)$ because $\inf_{\theta_i' \leq \theta_i} Q_i(\theta_i') + \eta(\theta_i - \theta_i') \leq Q_i(\theta_i)$ for all θ_i. For those types θ_i such that $U_i(\theta_i) < Q_i(\theta_i)$, by definition of U_i, there exists some $\theta_i' < \theta_i$ such that $U_i(\theta_i) = \min\{\underline{u}_i + \eta(\theta_i - \underline{\theta}_i),\ Q_i(\theta_i') + \eta(\theta_i - \theta_i')\}$, implying that $U_i'(\theta_i) = \eta$. For those types θ_i such that $U_i(\theta_i) = Q_i(\theta_i)$, by definition, $Q_i(\theta_i') + \eta(\theta_i - \theta_i') \geq Q_i(\theta_i)$ for all $\theta_i' < \theta_i$. This could happen only when $Q_i'(\theta_i) < \eta$, implying that $U_i'(\theta_i) < \eta$. Hence $(\boldsymbol{Q}, \boldsymbol{U})$ satisfies (IC) and is implementable by a contest.

Next we show that $\boldsymbol{U}$ is the unique utility profile such that $(\boldsymbol{Q}, \boldsymbol{U})$ is implementable by a contest, given utilities for the lowest types $\{\underline{u}_i\}_{i=1,\ldots,n}$. Suppose $\boldsymbol{U}^\dagger$ is a different utility profile such that $(\boldsymbol{Q}, \boldsymbol{U}^\dagger)$ is implementable by a contest and $U_i^\dagger(\underline{\theta}_i) = \underline{u}_i$ for all i. (IC) implies that $U_i^\dagger(\theta_i) \leq Q_i(\theta_i)$ for all θ_i.

Suppose there exists θ_i such that $U_i^\dagger(\theta_i) > U_i(\theta_i)$. This is only possible if $U_i(\theta_i) < Q_i(\theta_i)$ and there exists some $\theta_i' < \theta_i$ such that $U_i(\theta_i) = Q_i(\theta_i') + \eta(\theta_i - \theta_i') < U_i^\dagger(\theta_i)$. However, this implies that in the direct mechanism $(\boldsymbol{Q}, \boldsymbol{U}^\dagger)$, agent i with type θ_i' has incentives to misreport his type as θ_i. This contradicts $(\boldsymbol{Q}, \boldsymbol{U}^\dagger)$ is implementable by a contest.

Suppose there exists type θ_i such $U_i^\dagger(\theta_i) < U_i(\theta_i)$. This is only possible if $U_i^\dagger(\theta_i) < Q_i(\theta_i)$. Let $\theta_i' = \underline{\theta}_i$ if $U_i^\dagger(\theta_i) < Q_i(\theta_i)$ for all θ_i. Otherwise, let $\theta_i' = \sup\{z \leq \theta_i : U_i^\dagger(\theta_i') = Q_i(\theta_i')\}$. In both cases, by (IC), we have $U_i^\dagger(\theta_i) = U_i^\dagger(\theta_i') + \eta(\theta_i - \theta_i')$. In the case where $\theta_i' = \underline{\theta}_i$, we have

$$U_i^\dagger(\theta_i) = U_i^\dagger(\theta_i') + \eta(\theta_i - \theta_i') < U_i(\theta_i) \leq \underline{u}_i + \eta(\theta_i - \theta_i'),$$

implying that$U_i^\dagger(\underline{\theta}_i) < \underline{u}_i$, a contradiction. In the case where $\theta_i' > \underline{\theta}_i$, similarly we could infer $U_i^\dagger(\theta_i') < Q_i(\theta_i')$, again a contradiction. Hence, for any interim allocation rule $\boldsymbol{Q}$, if there exists interim utility $\boldsymbol{U}$ such that $(\boldsymbol{Q}, \boldsymbol{U})$ is implementable by a contest, then $\boldsymbol{U}$ is uniquely pinned down by $\boldsymbol{Q}$ and the utility profile for the lowest type, and its expression is given by (4).

Finally, for any $\boldsymbol{U}^\dagger$ such that $(\boldsymbol{Q}, \boldsymbol{U}^\dagger)$ is implementable by a contest, if $U_i^\dagger(\underline{\theta}_i) < U_i(\underline{\theta}_i)$ for all i, by (4), we must have $U_i^\dagger(\theta_i) \leq U_i(\theta_i)$ for all θ_i. □

Proof of Proposition 2. Notice that both "second-price" format mechanism and WTA contest implement the efficient allocation Q_{E}. Let the interim utility profiles under WTA contest and "second-price" format mechanism be $\boldsymbol{U}^{\mathcal{C}}$ and $\boldsymbol{U}^{\mathcal{S}}$ correspondingly.

By Corollary 1, $U_i^{\mathcal{C}}(\theta_i) \geq U_i^{\mathcal{S}}(\theta_i)$ for all θ_i and i. It remains to show that $U_i^{\mathcal{C}}(\theta_i) > U_i^{\mathcal{S}}(\theta_i)$ for some θ_i and i. By direct calculation,

$$\begin{aligned} U_i^{\mathcal{S}}(\theta_i) &= \eta(\theta_i - \mathbf{E}\left[\theta_{(k+1)} | \theta_i - \frac{1}{\eta} < \theta_{(k+1)} < \theta_i\right]) \times \mathbf{Pr}\left[\theta_i - \frac{1}{\eta} < \theta_{(k+1)} < \theta_i\right] + \mathbf{Pr}\left[\theta_i \geq \frac{1}{\eta} + \theta_{(k+1)}\right] \\ &< \mathbf{Pr}\big[\theta_i \geq \theta_{(k+1)}\big] = Q_{\mathrm{E}}(\theta_i). \end{aligned}$$

We distinguish two cases based on the properties of the distribution.

Case 1: Suppose the distribution F is such that $Q'_{\mathrm{E}}(\theta) \leq \eta$ for any $\theta \in (\underline{\theta}_i, \underline{\theta}_i + \epsilon)$ for any small $\epsilon > 0$. Then by Lemma 1, $U_i^{\mathcal{C}}(\theta_i) = Q_{\mathrm{E}}(\theta_i)$ for any $\theta \in (\underline{\theta}_i, \underline{\theta}_i + \epsilon)$. And since $Q_{\mathrm{E}}(\theta_i) > U_i^{\mathcal{S}}(\theta_i)$ for all θ_i, we have $U_i^{\mathcal{C}}(\theta_i) > U_i^{\mathcal{S}}(\theta_i)$ for any $\theta \in (\underline{\theta}_i, \underline{\theta}_i + \epsilon)$.

Case 2: Suppose the distribution F is such that $Q'_{\mathrm{E}}(\theta) > \eta$ for any $\theta \in (\underline{\theta}_i, \underline{\theta}_i + \epsilon)$ for any small $\epsilon > 0$. Then by Lemma 1, $U_i^{\mathcal{C}}(\theta_i)$ is linear with slope η for any $\theta \in (\underline{\theta}_i, \underline{\theta}_i + \epsilon)$. It remains to show that $U_i^{\mathcal{S}}(\theta_i)$ is not linear. Notice that

$$\begin{aligned} U_i^{\mathcal{S}}(\theta_i) &= \eta(\theta_i - \mathbf{E}\left[\theta_{(k+1)} | \theta_i - \frac{1}{\eta} < \theta_{(k+1)} < \theta_i\right]) \times \mathbf{Pr}\left[\theta_i - \frac{1}{\eta} < \theta_{(k+1)} < \theta_i\right] + \mathbf{Pr}\left[\theta_i \geq \frac{1}{\eta} + \theta_{(k+1)}\right] \\ &= \eta(\theta_i - \mathbf{E}\left[\theta_{(k+1)} | \theta_i - \frac{1}{\eta} < \theta_{(k+1)} < \theta_i\right]) \times [Q_{\mathrm{E}}(\theta_i) - Q_{\mathrm{E}}(\theta_i - \frac{1}{\eta})] + Q_{\mathrm{E}}(\theta_i - \frac{1}{\eta}) \\ &= \eta(\theta_i \times [Q_{\mathrm{E}}(\theta_i) - Q_{\mathrm{E}}(\theta_i - \frac{1}{\eta})] - \mathbf{E}\left[\theta_{(k+1)} \mathbf{1}\left[\theta_i - \frac{1}{\eta} < \theta_{(k+1)} < \theta_i\right]\right]) + Q_{\mathrm{E}}(\theta_i - \frac{1}{\eta}) \\ &= \eta \int_{\theta_i - \frac{1}{\eta}}^{\theta_i} Q_{\mathrm{E}}(\theta) d\theta, \end{aligned}$$

where the last equality is obtained by applying integration by part, i.e., $\mathbf{E}\left[\theta_{(k+1)} \mathbf{1}\left[\theta_i - \frac{1}{\eta} < \theta_{(k+1)} < \theta_i\right]\right] = \theta_i Q_{\mathrm{E}}(\theta_i) - (\theta_i - \frac{1}{\eta}) Q_{\mathrm{E}}(\theta_i - \frac{1}{\eta}) - \int_{\theta_i - \frac{1}{\eta}}^{\theta_i} Q_{\mathrm{E}}(\theta) d\theta$. Therefore, $U_i^{\mathcal{S}}(\theta_i)$ cannot be linear for a generic distribution. □

A.2 Proof omitted in Section 5

To simplify the notation in the later analysis, given the partition of the type space, we add a degenerate interval $\underline{\theta}^{(0)} = \bar{\theta}^{(0)} = \bar{\theta}$.

Proof of Theorem 2. By Lemma 1, the optimal utility function U_α must be continuous with sub-

gradient between 0 and η, and $Q_\alpha(\theta) = U_\alpha(\theta)$ if $Q'_\alpha(\theta) < \eta$. Therefore, we can partition the type space into countably many disjoint intervals $\{(\underline{\theta}^{(j)}, \bar{\theta}^{(j)})\}_{j=1}^{\infty}$ that fall into one of the following three categories:

Case 1: $Q_\alpha(\theta) = U_\alpha(\theta)$ and $U'_\alpha(\theta) < \eta$ for any type $\theta \in (\underline{\theta}^{(j)}, \bar{\theta}^{(j)})$;

Case 2: $Q_\alpha(\theta) = U_\alpha(\theta)$ and $U'_\alpha(\theta) = \eta$ for any type $\theta \in (\underline{\theta}^{(j)}, \bar{\theta}^{(j)})$;

Case 3: $Q_\alpha(\theta) > U_\alpha(\theta)$ and $U'_\alpha(\theta) = \eta$ for any type $\theta \in (\underline{\theta}^{(j)}, \bar{\theta}^{(j)})$;

For any interim allocation rule Q, define $\mathcal{Q}(\theta) = \int_\theta^{\bar{\theta}} Q(t)\,\mathrm{d}F(t)$. Notice that $\mathcal{Q}(\theta)$ is a continuous function.

Lemma 8. *If Q is optimal, then $\mathcal{Q}(\theta) - \int_\theta^{\bar{\theta}} Q_{\mathrm{E}}(t)\,\mathrm{d}F(t) < 0$ implies*

(A) $U(\theta) = Q(\theta)$; and

(B) either $U'(\theta) = \eta$ or $U'(\theta) = 0$.

Corollary 2. *If (Q, U) is optimal, then $Q(\theta) > U(\theta)$ implies $\mathcal{Q}(\theta) - \int_\theta^{\bar{\theta}} Q_{\mathrm{E}}(t)\,\mathrm{d}F(t) = 0$ and $Q(\theta) = Q_{\mathrm{E}}(\theta)$ a.e.*

Corollary 3. *If (Q, U) is optimal, then $0 < U'(\theta) < \eta$ implies $\mathcal{Q}(\theta) - \int_\theta^{\bar{\theta}} Q_{\mathrm{E}}(t)\,\mathrm{d}F(t) = 0$ and $Q(\theta) = Q_{\mathrm{E}}(\theta)$ a.e.*

Corollary 2 implies that for Case 3, we have $Q_\alpha(\theta) = Q_{\mathrm{E}}(\theta) > U_\alpha(\theta)$. Thus Case 3 will correspond to the efficient region.

The analysis of Case 1 is decomposed into two separate subcases.

Case 1a: $Q_\alpha(\theta) = U_\alpha(\theta)$ and $U'_\alpha(\theta) \in (0, \eta)$ for any type $\theta \in (\underline{\theta}^{(j)}, \bar{\theta}^{(j)})$;

Case 1b: $Q_\alpha(\theta) = U_\alpha(\theta)$ and $U'_\alpha(\theta) = 0$ for any type $\theta \in (\underline{\theta}^{(j)}, \bar{\theta}^{(j)})$.

By Corollary 3, in Case 1a, we have $\mathcal{Q}(\theta) - \int_\theta^{\bar{\theta}} Q_{\mathrm{E}}(t)\,\mathrm{d}F(t) = 0$ and $Q_\alpha(\theta) = Q_{\mathrm{E}}(\theta)$ for any type $\theta \in (\underline{\theta}^{(j)}, \bar{\theta}^{(j)})$. Therefore, Case 1a corresponds to the no-tension region. Moreover, we show that Case 1b cannot exist, with proofs deferred to the end of the section. Thus Case 1 is the no-tension region.

Lemma 9. *Case 1b does not exist in the optimal solution.*

Finally, for any interval j that corresponds to Case 2, if $\underline{\theta}^{(j)} > \underline{\theta}$, since the integration constraints ($\widehat{\mathrm{IF}}$) bind for all types within each interval given any of the two other cases, it must also bind for both of the end points of interval j, and hence

$$\int_{\underline{\theta}^{(j)}}^{\bar{\theta}^{(j)}} Q_\alpha(\theta)\,\mathrm{d}F(\theta) = \int_{\underline{\theta}^{(j)}}^{\bar{\theta}^{(j)}} Q_{\mathrm{E}}(\theta)\,\mathrm{d}F(\theta).$$

If $\underline{\theta}^{(j)} = \underline{\theta}$, the integration constraint ($\widehat{\text{IF}}$) also binds at $\underline{\theta}$ since otherwise we can increase the allocation and utility for a sufficiently small region above type $\underline{\theta}$ without violating the feasibility, which is a contradiction to the optimality of the solution. Hence the above equality holds as well, and Case 2 corresponds to the no-effort region. □

Proof of Lemma 8. Consider the following relaxed problem of ($\hat{\mathcal{P}}_\alpha$). We omit the monotonicity constraint on allocation as the relaxation of this constraint will not affect the optimal solution (Theorem 1).

$$\begin{aligned} \sup_{Q,U} \quad & \mathbf{E}_\theta[\alpha \cdot \theta \cdot Q(\theta) + (1-\alpha) \cdot U(\theta)] \\ \text{s.t.} \quad & \int_\theta^{\bar{\theta}} Q(\theta)\,\mathrm{d}F(z) \leq \int_\theta^{\bar{\theta}} Q_{\mathrm{E}}(z)\,\mathrm{d}F(z), \quad \forall \theta \in [\underline{\theta}, \bar{\theta}] \\ & U(\theta) \leq Q(\theta), 0 \leq U'(\theta) \leq a \end{aligned} \tag{$\hat{\mathcal{R}}_\alpha$}$$

Define $\mathcal{Q}(\theta) = \int_\theta^{\bar{\theta}} Q(t)\,\mathrm{d}F(t)$ and $\mathcal{Q}'(\theta) = -Q(\theta)f(\theta)$. The relaxed problem can be rewritten as

$$\begin{aligned} \sup_{\mathcal{Q},U} \quad & \int_{\underline{\theta}}^{\bar{\theta}} -\alpha \cdot \theta \cdot \mathcal{Q}'(\theta) + (1-\alpha) \cdot U(\theta) \cdot f(\theta)\,\mathrm{d}\theta \\ \text{s.t.} \quad & \mathcal{Q}(\theta) \leq \int_\theta^{\bar{\theta}} F^{n-1}(t)\,\mathrm{d}F(t) && \lambda(\theta) \\ & U(\theta)f(\theta) + \mathcal{Q}'(\theta) \leq 0 && \gamma(\theta) \\ & 0 \leq U'(\theta) && \kappa_1(\theta) \\ & U'(\theta) \leq a && \kappa_2(\theta) \end{aligned}$$

The Lagrange multipliers $\lambda(\theta), \gamma(\theta), \kappa_1(\theta), \kappa_2(\theta)$ are non-negative. The Lagrangian is given by

$$\begin{aligned} \hat{\mathcal{L}}(\mathcal{Q}, \mathcal{Q}', U, U', \lambda, \gamma, \kappa_1, \kappa_2) = & -[\alpha\theta \cdot \mathcal{Q}'(\theta) - (1-\alpha) \cdot U(\theta)f(\theta) \\ & + \lambda(\theta)(\mathcal{Q}(\theta) - \int_\theta^{\bar{\theta}} Q_{\mathrm{E}}(t)\,\mathrm{d}F(t)) \\ & + \gamma(\theta)(U(\theta)f(\theta) + \mathcal{Q}'(\theta)) \\ & + \kappa_1(\theta)(U'(\theta) - a) - \kappa_2(\theta)U'(\theta)] \end{aligned}$$

The solution of the problem satisfies the following conditions.

(1) The Euler-Lagrange conditions.[31]

$$\frac{\partial \hat{\mathcal{L}}}{\partial \mathcal{Q}} - \frac{\mathrm{d}}{\mathrm{d}\theta}\frac{\partial \hat{\mathcal{L}}}{\partial \mathcal{Q}'} = 0 \Leftrightarrow \quad \lambda(\theta) - (\alpha + \gamma'(\theta)) = 0 \tag{E-L-1}$$

and

$$\frac{\partial \hat{\mathcal{L}}}{\partial U} - \frac{\mathrm{d}}{\mathrm{d}\theta}\frac{\partial \hat{\mathcal{L}}}{\partial U'} = 0 \Leftrightarrow \quad (\gamma(\theta) - (1-\alpha))f(\theta) - \kappa'(\theta) = 0 \tag{E-L-2}$$

hold whenever well-defined, where $\kappa(\theta) = \kappa_1(\theta) - \kappa_2(\theta)$.

(2) The complementary slackness conditions

$$\lambda(\theta)(\mathcal{Q}(\theta) - \int_{\theta}^{\bar{\theta}} F^{n-1}(t)\,\mathrm{d}F(t)) = 0, \quad \lambda(\theta) \geq 0 \tag{C-S-1a}$$

$$\gamma(\theta)[U(\theta)f(\theta) + \mathcal{Q}'(\theta)] = 0, \quad \gamma(\theta) \geq 0 \tag{C-S-1b}$$

$$\kappa_1(\theta)[U'(\theta) - a] = 0, \quad \kappa_1(\theta) \geq 0 \tag{C-S-1c}$$

$$\kappa_2(\theta)U'(\theta) = 0, \quad \kappa_2(\theta) \geq 0 \tag{C-S-1d}$$

Suppose $\mathcal{Q}(\theta) - \int_{\theta}^{\bar{\theta}} Q_{\mathrm{E}}(t)\,\mathrm{d}F(t) < 0$. We show that the following two conditions hold for the optimal solution.

- $U(\theta) = Q(\theta)$. By (C-S-1a), $\lambda(\theta) = 0$ holds in an interval. From (E-L-1), we have $\gamma'(\theta) = -\alpha$. Hence $\gamma(\theta)$ cannot be a constant in this interval, in particular, $\gamma(\theta) \neq 0$ except for at most one point. Combined with (C-S-1b) further implies $U(\theta)f(\theta) + \mathcal{Q}'(\theta) = 0$, i.e., $U(\theta) = Q(\theta)$.

- Either $U'(\theta) = a$ or $U'(\theta) = 0$. Similarly, $\gamma(\theta) \neq 1 - \alpha$ except for at most one point, and combined with (E-L-2) implies that $\kappa'(\theta) \neq 0$ except for at most one point. This suggests that $\kappa(\theta)$ is not a constant, in particular, not zero. Then the result is implied by applying (C-S-1c) and (C-S-1d). □

Proof of Corollary 2. The contra-positive of Lemma 8 is also true: $Q(\theta) > U(\theta)$ implies $\mathcal{Q}(\theta) - \int_{\theta}^{\bar{\theta}} Q_{\mathrm{E}}(t)\,\mathrm{d}F(t) = 0$. By rearranging the terms and taking the derivative over θ, we have $Q(\theta) = Q_{\mathrm{E}}(\theta)$ almost everywhere. □

[31] We are looking for piece-wise continuous solutions (the state variables are continuous and the control variables are piece-wise continuous), since in principle, the allocation $Q(\theta)$ does not have to be continuous; instead, it could be piece-wise continuous, while $U(\theta)$ is continuous but its derivative might not be. The necessary conditions should be the integral form of Euler-Lagrange conditions, and the Erdmann-Weierstrass corner conditions (c.f., Clarke, 2013). However, the Erdmann-Weierstrass corner conditions have no bite here. And we use the usual form of Euler-Lagrange conditions, since it does not involve the state variables or the controls. Notice, though, the Lagrange multiplier $\gamma(\theta)$ could potentially be PC^1.

Proof of Lemma 9. Suppose Case 1b exists. In this case, since U is a continuous function, $Q_\alpha(\theta) = U_\alpha(\theta) = U_\alpha(\bar{\theta}^{(j)})$ for any type $\theta \in (\underline{\theta}^{(j)}, \bar{\theta}^{(j)})$. Let j' be the index of the interval such that $\bar{\theta}^{(j)} = \underline{\theta}^{(j')}$. We consider three possible situations for interval j'.

- Interval for j' belongs to Case 1. In this case, the integration constraint ($\widehat{\text{IF}}$) bind at $\underline{\theta}^{(j')}$ and $U(\bar{\theta}^{(j)}) = U(\underline{\theta}^{(j')}) = Q_{\text{E}}(\underline{\theta}^{(j')})$. Therefore, there exists a sufficiently small constant $\epsilon > 0$ such that the integration constraint ($\widehat{\text{IF}}$) is violated at type $\bar{\theta}^{(j)} - \epsilon$, a contradiction.

- Interval for j' belongs to Case 2. In this case, integration constraint ($\widehat{\text{IF}}$) does not bind at type $\underline{\theta}^{(j')}$. Suppose otherwise, we must have $Q_{\text{E}}(\underline{\theta}^{(j')}) \leq U(\bar{\theta}^{(j)})$ in order for the integration constraint ($\widehat{\text{IF}}$) to hold for type $\underline{\theta}^{(j')} + \epsilon$ given sufficiently small $\epsilon > 0$. However, this would imply that the integration constraint ($\widehat{\text{IF}}$) is violated for type $\bar{\theta}^{(j)} - \epsilon$ given sufficiently small $\epsilon > 0$.

  Next we consider two cases for interval j.

  – $\underline{\theta}^{(j)} > \underline{\theta}$. In this case, the integration constraint ($\widehat{\text{IF}}$) cannot bind at any type $\theta \in [\underline{\theta}^{(j)}, \bar{\theta}^{(j)})$. This is because if it binds at θ, it implies that $Q(\theta) = U(\theta) > Q_{\text{E}}(\theta)$. By the continuity of U and Q_{E}, and the constraint that $Q \geq U$, there exists a sufficiently small constant $\epsilon > 0$ such that the integration constraint ($\widehat{\text{IF}}$) is violated at $\theta - \epsilon$. Thus, there exists $\epsilon, \delta > 0$ such that for any type $\theta \in [\underline{\theta}^{(j)} - \epsilon, \bar{\theta}^{(j)} + \epsilon]$,

$$\int_\theta^{\bar{\theta}} Q(z)\,\mathrm{d}F(z) \leq \int_\theta^{\bar{\theta}} Q_{\text{E}}(z)\,\mathrm{d}F(z) - \delta.$$

  Moreover, we select ϵ to be sufficient small to satisfy the additional condition that $Q'(\theta) \leq \eta$ for any type $\theta \in [\underline{\theta}^{(j)} - \epsilon, \bar{\theta}^{(j)} + \epsilon]$. Given parameter θ^*, let $Q^\ddagger$ be the allocation such that

  1. $Q^\ddagger(\theta) = Q(\underline{\theta}^{(j)} - \epsilon)$ for any type $\theta \in [\underline{\theta}^{(j)} - \epsilon, \theta^*]$;
  2. $Q^\ddagger(\theta) = Q(\underline{\theta}^{(j)} - \epsilon) + \eta \cdot (\theta - \theta^*)$ for any type $\theta \in (\theta^*, \theta^* + \frac{1}{\eta} \cdot Q(\bar{\theta}^{(j)} + \epsilon) - Q(\underline{\theta}^{(j)} - \epsilon))$;
  3. $Q^\ddagger(\theta) = Q(\bar{\theta}^{(j)} + \epsilon)$ for any type $\theta \in [\theta^* + \frac{1}{\eta} \cdot Q(\bar{\theta}^{(j)} + \epsilon) - Q(\underline{\theta}^{(j)} - \epsilon), \bar{\theta}^{(j)} + \epsilon]$.

  The parameter θ^* is chosen such that

$$\int_{\underline{\theta}^{(j)} - \epsilon}^{\bar{\theta}^{(j)} + \epsilon} Q^\ddagger(z)\,\mathrm{d}F(z) = \int_{\underline{\theta}^{(j)} - \epsilon}^{\bar{\theta}^{(j)} + \epsilon} Q(z)\,\mathrm{d}F(z).$$

  It is easy to verify that

$$\int_{\underline{\theta}^{(j)} - \epsilon}^{\bar{\theta}^{(j)} + \epsilon} z \cdot Q^\ddagger(z)\,\mathrm{d}F(z) > \int_{\underline{\theta}^{(j)} - \epsilon}^{\bar{\theta}^{(j)} + \epsilon} z \cdot Q(z)\,\mathrm{d}F(z)$$

since $Q^\ddagger$ shifts allocation probabilities from low types to high types compared to Q. Therefore, given a sufficiently small constant $\hat{\delta} > 0$, consider another allocation-utility pair $(Q^\dagger, U^\dagger)$ such that

1. $Q^\dagger(\theta) = Q(\theta)$ and $U^\dagger(\theta) = U(\theta)$ for any type $\theta \notin [\underline{\theta}^{(j)} - \epsilon, \bar{\theta}^{(j)} + \epsilon]$;
2. $Q^\dagger(\theta) = (1-\hat{\delta}) \cdot Q(\theta) + \hat{\delta} \cdot Q^\ddagger(\theta)$ and $U^\dagger(\theta) = (1-\hat{\delta}) \cdot U(\theta) + \hat{\delta} \cdot Q^\ddagger(\theta)$ for any type $\theta \in [\underline{\theta}^{(j)} - \epsilon, \bar{\theta}^{(j)} + \epsilon]$.

The new allocation-utility pair $(Q^\dagger, U^\dagger)$ is feasible and strictly improves the objective value, a contradiction to the optimality of (Q, U).

– $\underline{\theta}^{(j)} = \underline{\theta}$. The proof for this case is similar. The only difference is that we can change the allocation and utility within interval j without worrying about the continuity of utility function from lower types. Therefore, by adopting a similar construction for $Q^\ddagger$ and $(Q^\dagger, U^\dagger)$, restricted to the interval $[\underline{\theta}^{(j)}, \bar{\theta}^{(j)} + \epsilon]$ for sufficiently small $\epsilon > 0$, we can again show that the allocation-utility pair (Q, U) that contains Case 1b is not optimal.

- Either interval for j' belongs to Case 3, or $\underline{\theta}^{(j')}$ is the highest possible type $\bar{\theta}$. In either cases, both the efficient allocation Q_{E} and the interim allocation Q are strictly above the utility at $\underline{\theta}^{(j')}$ in order for the integration constraint $(\widehat{\mathrm{IF}})$ to be satisfied within interval j. Therefore, the allocation within interval j can be increased without violating the monotonicity compared to allocations above $\underline{\theta}^{(j')}$. Again we adopt a similar construction for $Q^\ddagger$ and $(Q^\dagger, U^\dagger)$, restricted to the interval $[\underline{\theta}^{(j)} - \epsilon, \bar{\theta}^{(j)}]$ for sufficiently small $\epsilon > 0$. Here we add the additional operation to increase the utility $U^\dagger$ for types above $\underline{\theta}^{(j')}$ to satisfy the monotonicity of the utility function, which only improves the objective value. Thus, we show that the allocation-utility pair (Q, U) that contains Case 1b is not optimal. □

A.3 Proof omitted in Section 7

Proof of Proposition 4. By Theorem 2, there exists a partition of the type space $\{(\underline{\theta}^{(j)}, \bar{\theta}^{(j)})\}_{j=1}^{\infty}$ such that each interval belongs to one the three cases. It is sufficient to show that the order of the three cases on the type space can not be changed in the optimal contest.

First we show that for j such that interval j is the no-tension region, then it is optimal for all intervals with types below $\underline{\theta}^{(j)}$ to be the no-tension region. The main reason is that by the convexity of the efficient allocation rule, for any type $\theta \leq \underline{\theta}^{(j)}$, $Q'_{\mathrm{E}}(\theta) \leq Q'_{\mathrm{E}}(\underline{\theta}^{(j)}) \leq \eta$. Therefore, by setting $U_\alpha(\theta) = Q_\alpha(\theta) = Q'_{\mathrm{E}}(\theta)$, the resulting contest is feasible and trivially maximize the objective value.

Let $\theta^{(1)}$ be the supreme of types θ such that θ is in the no-tension region. The argument in previous paragraph shows that the whole interval $(\underline{\theta}, \theta^{(1)})$ is the no-tension region. Moreover, according to Theorem 2, $Q_\alpha(\theta^{(1)}) = U_\alpha(\theta^{(1)}) = Q_{\mathrm{E}}(\theta^{(1)})$ and for any $\theta \geq \theta^{(1)}$, we have $U'_\alpha(\theta) = \eta$. Next we consider two different cases.

- If $Q'_{\mathrm{E}}(\theta^{(1)}) \geq \eta$, by convexity of the efficient allocation rule, $Q_{\mathrm{E}}(\theta) > U_\alpha(\theta)$ for any type $\theta > \theta^{(1)}$, which implies that

$$\int_{\underline{\theta}^{(j)}}^{\bar{\theta}^{(j)}} U_\alpha(\theta)\,\mathrm{d}F(\theta) < \int_{\underline{\theta}^{(j)}}^{\bar{\theta}^{(j)}} Q_{\mathrm{E}}(\theta)\,\mathrm{d}F(\theta).$$

  for any interval j with types above $\theta^{(1)}$. In this case, $\theta^{(1)} = \theta^{(2)}$ and the no-effort region does not exist.

- If $Q'_{\mathrm{E}}(\theta^{(1)}) < \eta$, $Q_{\mathrm{E}}(\theta) < U_\alpha(\theta)$ for type θ sufficiently close to $\theta^{(1)}$. Therefore, for interval j such that $\underline{\theta}^{(j)} = \theta^{(1)}$, interval j must be the no-effort region. Let $\theta^{(2)} = \bar{\theta}^{(j)}$. Note that in order for the integration constraint to be satisfied in interval j, we have $Q_{\mathrm{E}}(\theta^{(2)}) \geq U_\alpha(\theta^{(2)})$ and $Q'_{\mathrm{E}}(\theta^{(2)}) \geq \eta$. Therefore, for any type $\theta > \theta^{(2)}$, $Q_{\mathrm{E}}(\theta) > U_\alpha(\theta)$ and hence any interval above type $\theta^{(2)}$ is the efficient region. □

Proof of Lemma 6. Let n be a sufficiently large number such that $Q_{\mathrm{E},n}$ is convex, and let $\theta_n^\dagger$ be the cutoff type such that in the IC implementation of efficient allocation, agents exert costly effort with any type $\theta > \theta_n^\dagger$, i.e., $Q'_{\mathrm{E},n}(\theta_n^\dagger) = \eta$. That is, $(n-1)\cdot F^{n-2}(\theta_n^\dagger)\cdot f(\theta_n^\dagger) = \eta$. Rearranging the terms, we have

$$F^{n-2}(\theta_n^\dagger) = \frac{\eta}{(n-1)\cdot f(\theta_n^\dagger)}.$$

Note that by Assumption 2, the right hand side is lower bounded by $\frac{\eta}{(n-1)\cdot\bar{\beta}_1}$. Therefore, for any $\epsilon_0 > 0$, there exists N_0 such that for any $n \geq N_0$, we have

$$F(\theta_n^\dagger) \geq \left(\frac{\eta}{(n-1)\cdot\bar{\beta}_1}\right)^{\frac{1}{n-2}} \geq 1-\epsilon_0.$$

Since the density is lower bounded by $\underline{\beta}_1$, we have that $\theta_n^\dagger \geq \bar{\theta} - \frac{\epsilon_0}{\underline{\beta}_1}$. For any $\epsilon_1 > 0$, let N_1 be the integer such that $\frac{\eta}{(n-1)\cdot f(\theta_n^\dagger)} \leq \epsilon_1$ for any $n \geq N_1$. The expected utility of the agent with type $\bar{\theta}$ is

$$U_{\mathrm{E},n}(\bar{\theta}) = F^{n-1}(\theta_n^\dagger) + \eta(\bar{\theta} - \theta_n^\dagger) \leq F^{n-2}(\theta_n^\dagger) + \frac{\eta\cdot\epsilon_0}{\underline{\beta}_1} \leq \epsilon_1 + \frac{\eta\cdot\epsilon_0}{\underline{\beta}_1}.$$

Let $\theta_n^\ddagger$ be the type such that $F(\theta_n^\ddagger) = 1 - \frac{1}{n}$. There exists N_2 such that $\theta_n^\ddagger \geq \theta_n^\dagger$ for any $n \geq N_2$. For any $\epsilon > 0$, let $\epsilon_1 = \frac{\epsilon}{2}$, $\epsilon_0 = \frac{\epsilon\underline{\beta}_1}{2\eta}$, and $N = \max\{N_0, N_1, N_2\}$, for any $n \geq N$, the expected effort of any agent is at least his effort from types above $\theta_n^\ddagger$, which is lower bounded by

$$(1-F(\theta_n^\ddagger))\cdot(Q_{\mathrm{E},n}(\theta_n^\ddagger) - U_{\mathrm{E},n}(\bar{\theta})) \geq \frac{1}{n}\left(\frac{1}{e} - \epsilon_1 + \frac{\eta\cdot\epsilon_0}{\underline{\beta}_1}\right) = \frac{1}{n}\left(\frac{1}{e} - \epsilon\right).$$

Since the item is always allocated in equilibrium, the total utility is

$$n \cdot \mathbf{E}_{\theta\sim F}[U_{\mathrm{E},n}(\theta)] \le 1 - \frac{1}{e} + \epsilon. \qquad \square$$

Proof of Lemma 7. Note that compared to efficient allocation $Q_{\mathrm{E},n}$, the chosen allocation rule Q_n only randomizes the allocation for types between $\hat{\theta}_n$ and $\bar{\theta}$. Therefore, we have

$$n \cdot \mathbf{E}_{\theta\sim F}[\theta \cdot Q_n(\theta)] \ge n \cdot \mathbf{E}_{\theta\sim F}[\theta \cdot Q_{\mathrm{E},n}(\theta)] - (\bar{\theta} - \hat{\theta}_n).$$

Similar to the proof of Theorem 3, we can show that $\lim_{n\to\infty} \hat{\theta}_n = \bar{\theta}$. By taking the limit of the above inequality, we have that

$$\lim_{n\to\infty} n \cdot \mathbf{E}_{\theta\sim F}[\theta \cdot Q_n(\theta)] \ge \lim_{n\to\infty} n \cdot \mathbf{E}_{\theta\sim F}[\theta \cdot Q_{\mathrm{E},n}(\theta)] = \bar{\theta}.$$

Thus for any $\epsilon > 0$, there exists N_1 such that for any $n \ge N_1$, $n \cdot \mathbf{E}_{\theta\sim F}[\theta \cdot Q_n(\theta)] \ge \bar{\theta} - \epsilon$. $\square$

A.4 Proof of Theorem 5

It is tempting to conjecture that when z is large enough, $Q_{\mathrm{E},z}(\theta)$ has an S shape, i.e., convex for small θ and concave for large θ, which would naturally imply the order of the intervals as stated in our result. However, this in general is not true.[32] To circumvent this inconvenience, note that for any small constant ϵ_0, when z is large enough, the interim efficient allocation has small slope (smaller than the marginal cost of effort η) outside the small interval $(\theta_c - \epsilon_0, \theta_c + \epsilon_0)$ centered at θ_c. Moreover, since the value of the efficient allocation changes a lot in this small interval, the agents will exert high efforts in equilibrium if the items are allocated efficiently, leading to low expected utility for types around θ_c. We show that in the optimal contest, the principal randomizes the allocation around θ_c. In particular, the no-effort region where the allocation is randomized will cover the whole interval of $(\theta_c - \epsilon_0, \theta_c + \epsilon_0)$. Since the derivatives of the efficient allocation outside this region is at most η, the principal's objective value is maximized by allocating the items efficiently. Next we provide the formal proof.

Proof of Theorem 5. Since the distribution is continuous, the probability there is a tie for the types

[32] The second order derivative of the allocation is

$$Q''_{\mathrm{E},z}(\theta) = (zn-1) \cdot \binom{zn-2}{zk-1} (1-F(\theta))^{zk-2} \cdot (F(\theta))^{z(n-k)-2} \cdot \left(f^2(\theta)(z(n-k)-1-(zn-2)F(\theta)) + f'(\theta)(1-F(\theta))F(\theta)\right).$$

No matter how large the parameter z is, for types within $(\theta_c - \epsilon_0, \theta_c + \epsilon_0)$, the sign of the second order derivative could alternate multiple times.

is 0. Therefore, given scale parameter z, the interim efficient allocation is

$$Q_{\mathrm{E},z}(\theta) = \mathbf{Pr}\big[\theta_{(nz-kz:nz-1)} \le \theta\big] = \sum_{j=0}^{zk-1} \binom{zn-1}{j} \cdot (1-F(\theta))^j \cdot (F(\theta))^{zn-1-j}$$

where $\theta_{(nz-kz:nz-1)}$ is the $(nz-kz)$-th order statistics, i.e., the $(nz-kz)$-th smallest value in a sample of $nz-1$ observations, and the binomial coefficient $\binom{n}{k}$ is defined by the expression $\binom{n}{k} = \frac{n!}{k!(n-k)!}$

Recall that θ_c is the cutoff type such that $1-F(\theta_c) = \frac{k}{n}$. The derivative of the allocation is

$$Q'_{\mathrm{E},z}(\theta) = f(\theta) \cdot (zn-1) \cdot \binom{zn-2}{zk-1} (1-F(\theta))^{zk-1} \cdot (F(\theta))^{z(n-k)-1}.$$

Note that $\binom{zn-2}{zk-1}(1-F(\theta))^{zk-1} \cdot (F(\theta))^{z(n-k)-1}$ is the probability that the binomial random variable $B(zn-2, 1-F(\theta))$ equals $zk-1$. When $1-F(\theta) < \frac{k}{n}$, this probability is exponentially small as zn increases, which implies that $\lim_{z\to\infty} Q'_{\mathrm{E},z}(\theta) = 0$. Therefore, for any $\epsilon_0 > 0$, there exists Z_0 such that for any $z \ge Z_0$, for any type $\theta \notin [\theta_c - \epsilon_0, \theta_c + \epsilon_0]$,

$$Q'_{\mathrm{E},z}(\theta) \le \eta.$$

Again by Hoeffding's inequality, for any $\epsilon_1 > 0$, there exists Z_1 such that for any $z \ge Z_1$,

$$Q_{\mathrm{E},z}(\theta) \le \epsilon_1$$

for any type $\theta \le \theta_c - \epsilon_0$ and

$$Q_{\mathrm{E},z}(\theta) \ge 1 - \epsilon_1$$

for any type $\theta \ge \theta_c + \epsilon_0$. Intuitively, this is because $\lim_{z\to\infty} Q_{\mathrm{E},z}(\theta)$ is a step function, i.e.,

$$\lim_{z\to\infty} Q_{\mathrm{E},z}(\theta) = \begin{cases} 0 & \text{if } \theta < \theta_c \\ 1 & \text{if } \theta \ge \theta_c. \end{cases}$$

Let $\tilde{\theta}^{(1)} \triangleq \theta_c - \epsilon_0 - \sqrt{\frac{8\epsilon_0 \bar{\beta}_1}{\eta \underline{\beta}_{-1}}}$.

Lemma 10. *For sufficiently large z, in optimal contest $(Q_{\alpha,z}, U_{\alpha,z})$, we have $U_{\alpha,z}(\tilde{\theta}^{(1)}) > Q_{\mathrm{E},z}(\tilde{\theta}^{(1)})$.*

We defer the proof of the lemma to the end of the section. Note that in optimal contest $(Q_{\alpha,z}, U_{\alpha,z})$, $U_{\alpha,z}(\tilde{\theta}^{(1)}) > Q_{\mathrm{E},z}(\tilde{\theta}^{(1)})$ implies that type $\tilde{\theta}^{(1)}$ must belong to a no-effort region. Let $\theta^{(1)} < \tilde{\theta}^{(1)} < \theta^{(2)}$ be the end points of this no-effort region. Let Θ_+ be the set of types in $(\theta^{(1)}, \theta^{(2)})$ such that $Q_{\mathrm{E},z}(\theta) > \hat{Q}_{\alpha,z}(\theta)$ and Θ_- be the set of types in $(\theta^{(1)}, \theta^{(2)})$ such that $Q_{\mathrm{E},z}(\theta) < Q_{\alpha,z}(\theta)$.

Since the integration constraint binds within $(\theta^{(1)}, \theta^{(2)})$, we have that

$$\int_{\Theta_+} (Q_{\mathrm{E},z}(\theta) - Q_{\alpha,z}(\theta))\,\mathrm{d}F(\theta) + \int_{\Theta_-} (Q_{\mathrm{E},z}(\theta) - Q_{\alpha,z}(\theta))\,\mathrm{d}F(\theta) = 0.$$

Note that

$$\begin{aligned}\int_{\Theta_-} (Q_{\mathrm{E},z}(\theta) - Q_{\alpha,z}(\theta))\,\mathrm{d}F(\theta) &\leq -\int_{\theta^{(1)}}^{\theta_c-\epsilon_0} (\eta(\theta-\theta^{(1)}) - \epsilon_1)\,\mathrm{d}F(\theta) \\ &\leq -\underline{\beta}_1 \cdot \left(\frac{\eta}{2}\cdot(\theta_c-\epsilon_0-\theta^{(1)})^2 - \epsilon_1\cdot(\theta_c-\epsilon_0-\theta^{(1)})\right).\end{aligned}$$

Similarly,

$$\int_{\Theta_+} (Q_{\mathrm{E},z}(\theta) - Q_{\alpha,z}(\theta))\,\mathrm{d}F(\theta) \leq \int_{\theta_c-\epsilon_0}^{\theta^{(2)}} 1\,\mathrm{d}F(\theta) \leq \bar{\beta}_1\cdot(\theta^{(2)}-\theta_c+\epsilon_0).$$

Combining the inequalities above, for sufficiently small $\epsilon_1 \leq \frac{\eta}{4}(\theta_c-\epsilon_0-\theta^{(1)})$, we must have

$$\theta^{(2)} \geq \theta_c - \epsilon_0 + \frac{\eta\cdot\underline{\beta}_1}{4\bar{\beta}_1}\cdot(\theta_c-\epsilon_0-\theta^{(1)})^2 \geq \theta_c+\epsilon_0.$$

The last inequality holds by simply substituting the bound for $\theta^{(1)}$. This implies that in the optimal contest, the no-effort region $(\theta^{(1)}, \theta^{(2)})$ covers the whole interval of $(\theta_c-\epsilon_0, \theta_c+\epsilon_0)$. Note that since the derivatives of the efficient allocation outside the no-effort region $(\theta^{(1)}, \theta^{(2)})$ is at most η, the objective of the principal is maximized by allocating the items efficiently. In particular, letting $\theta^{(3)} \geq \theta^{(2)}$ be the type such that the linear extension of the utility function within the no-tension region intersects with the efficient allocation rule. The interval $(\theta^{(2)}, \theta^{(3)})$ is the efficient region, and $(\underline{\theta}, \theta^{(1)})$ and $(\theta^{(3)}, \bar{\theta})$ are the no-tension regions. □

Proof of Lemma 10. It is sufficient to show that any contest $(\tilde{Q}_{\alpha,z}, \tilde{U}_{\alpha,z})$ such that $\tilde{U}_{\alpha,z}(\tilde{\theta}^{(1)}) \leq Q_{\mathrm{E},z}(\tilde{\theta}^{(1)})$ cannot be an optimal contest. We prove by contradiction. In what follows, we construct a contest $\hat{Q}_{\alpha,z}, \hat{U}_{\alpha,z}$ and show that it achieves higher objective value.

Let $\epsilon_0, \epsilon_1, \epsilon_2 > 0$ be any number such that[33]

$$0 < \epsilon_0 \le \min\left\{\frac{\underline{\beta}_1}{10\eta\cdot\bar{\beta}_1}, \epsilon_2^4\right\}, \qquad \epsilon_0 + 2\sqrt{\frac{8\epsilon_0\bar{\beta}_1}{\eta\underline{\beta}_1}} \le \epsilon_2,$$

$$0 < \epsilon_1 \le \min\{0.01, \epsilon_2^4\}, \qquad 0 < \epsilon_2 < \frac{\underline{\beta}_1}{10\eta\cdot\bar{\beta}_1},$$

$$\alpha\bar{\beta}_1 \cdot \left((\epsilon_2+\epsilon_0)^2\cdot\frac{\bar{\beta}_1}{\underline{\beta}_1} + \epsilon_0 + \epsilon_2\right)^2 < \frac{1}{2\eta}(1-\alpha)\cdot\left(\eta\left(\epsilon_2 - \epsilon_0 - \sqrt{\frac{8\epsilon_0\bar{\beta}_1}{\eta\underline{\beta}_1}}\right) - \epsilon_1\right).$$

Let $\hat{\theta}^{(1)} \triangleq \theta_c - \epsilon_2$. By the choice of ϵ_0, we have $\hat{\theta}^{(1)} < \tilde{\theta}^{(1)}$.

Consider a contest $(\hat{Q}_{\alpha,z}, \hat{U}_{\alpha,z})$ characterized by three cutoffs $\hat{\theta}^{(1)} < \hat{\theta}^{(2)} \le \hat{\theta}^{(3)}$ such that $(\underline{\theta}, \hat{\theta}^{(1)})$ is the no-tension region, $(\hat{\theta}^{(1)}, \hat{\theta}^{(2)})$ is the no-effort region, $(\hat{\theta}^{(2)}, \hat{\theta}^{(3)})$ is the efficient region, and $(\hat{\theta}^{(3)}, \bar{\theta})$ is the no-tension region.

Step 1: In this step, we will show that if $\hat{\theta}^{(1)}$ is chosen such that $\theta_c - \frac{\underline{\beta}_1}{10\eta\cdot\bar{\beta}_1} \le \hat{\theta}^{(1)}$,[34] then the integration constraint for the no-effort interval imposes an upper bound on the length of the no-effort interval, i.e., $\hat{\theta}^{(2)} \le \tilde{\theta}$, where $\tilde{\theta} \triangleq \theta_c + \epsilon_0 + \frac{2\eta\cdot\bar{\beta}_1}{\underline{\beta}_1}\cdot(\theta_c + \epsilon_0 - \hat{\theta}^{(1)})^2$.

Let $\hat{\Theta}_+$ be the set of types in $(\hat{\theta}^{(1)}, \hat{\theta}^{(2)})$ such that $Q_{\mathrm{E},z}(\theta) > \hat{Q}_{\alpha,z}(\theta)$ and $\hat{\Theta}_-$ be the set of types in $(\hat{\theta}^{(1)}, \hat{\theta}^{(2)})$ such that $Q_{\mathrm{E},z}(\theta) < \hat{Q}_{\alpha,z}(\theta)$. Since the integration constraint binds within $(\hat{\theta}^{(1)}, \hat{\theta}^{(2)})$, we have that

$$\begin{aligned} 0 &= \int_{\hat{\Theta}_+} (Q_{\mathrm{E},z}(\theta) - \hat{Q}_{\alpha,z}(\theta))\,\mathrm{d}F(\theta) + \int_{\hat{\Theta}_-} (Q_{\mathrm{E},z}(\theta) - \hat{Q}_{\alpha,z}(\theta))\,\mathrm{d}F(\theta) \\ &\ge \int_{\theta_c+\epsilon_0}^{\hat{\theta}^{(2)}} (1 - 2\epsilon_1 - \eta(\theta - \hat{\theta}^{(1)}))\,\mathrm{d}F(\theta) - \int_{\hat{\theta}^{(1)}}^{\theta_c+\epsilon_0} \eta(\theta - \hat{\theta}^{(1)})\,\mathrm{d}F(\theta). \end{aligned}$$

By the choice of $\hat{\theta}^{(1)}$ and ϵ_0, ϵ_1, $1 - 2\epsilon_1 - \eta(\theta - \hat{\theta}^{(1)}) \ge \frac{1}{2}$ for any type $\theta \le \tilde{\theta}$. Therefore,

$$\begin{aligned} &\int_{\theta_c+\epsilon_0}^{\tilde{\theta}} (1 - 2\epsilon_1 - \eta(\theta - \hat{\theta}^{(1)}))\,\mathrm{d}F(\theta) - \int_{\hat{\theta}^{(1)}}^{\theta_c+\epsilon_0} \eta(\theta - \hat{\theta}^{(1)})\,\mathrm{d}F(\theta) \\ &\ge \frac{\underline{\beta}_1}{2}(\tilde{\theta} - \theta_c - \epsilon_0) - \eta\cdot\bar{\beta}_1(\theta_c + \epsilon_0 - \hat{\theta}^{(1)})^2 \ge 0. \end{aligned}$$

Combining the above two inequalities, we get the desired bound on $\hat{\theta}^{(2)}$.

Step 2: Next we utilize the upper bound to show that the objective value of the contest $\hat{Q}_{\alpha,z}, \hat{U}_{\alpha,z}$ is higher than that of contest $\tilde{Q}_{\alpha,z}, \tilde{U}_{\alpha,z}$ with $\tilde{U}_{\alpha,z}(\tilde{\theta}^{(1)}) \le \tilde{Q}_{\alpha,z}(\tilde{\theta}^{(1)})$. Note that $Q_{\mathrm{E},z}$ and $\hat{Q}_{\alpha,z}(\theta)$

[33] Notice that these inequalities could hold at the same time: one could choose ϵ_0 and ϵ_1 that are "small" compared to ϵ_2, for example, $\epsilon_0 = o(\epsilon_2^4)$ and $\epsilon_1 = o(\epsilon_2^4)$, then the last inequality holds because the left hand side is of higher order than the right hand side.

[34] Such a choice is possible because by the choice of $\epsilon_0, \epsilon_1, \epsilon_2$, we have $\theta_c - \frac{\underline{\beta}_1}{10\eta\cdot\bar{\beta}_1} \le \tilde{\theta}^{(1)}$.

coincides at any type θ outside the no-effort region. Therefore, the loss in efficiency compared to the efficient allocation rule is

$$\begin{aligned}
&\alpha \cdot \int_{\hat{\theta}^{(1)}}^{\hat{\theta}^{(2)}} \theta \cdot Q_{\mathrm{E},z}\,\mathrm{d}F(\theta) - \alpha \cdot \int_{\hat{\theta}^{(1)}}^{\hat{\theta}^{(2)}} \theta \cdot \hat{Q}_{\alpha,z}(\theta)\,\mathrm{d}F(\theta) \\
=&\,\alpha \cdot \int_{\hat{\Theta}_+} \theta \cdot (Q_{\mathrm{E},z}(\theta) - \hat{Q}_{\alpha,z}(\theta))\,\mathrm{d}F(\theta) - \alpha \cdot \int_{\hat{\Theta}_-} \theta \cdot (Q_{\mathrm{E},z}(\theta) - \hat{Q}_{\alpha,z}(\theta))\,\mathrm{d}F(\theta) \\
\leq&\,\alpha \cdot (\hat{\theta}^{(2)} - \hat{\theta}^{(1)}) \cdot \int_{\hat{\Theta}_+} (Q_{\mathrm{E},z}(\theta) - \hat{Q}_{\alpha,z}(\theta))\,\mathrm{d}F(\theta) \\
\leq&\,\alpha \cdot (\hat{\theta}^{(2)} - \hat{\theta}^{(1)}) \cdot (F(\hat{\theta}^{(2)}) - F(\hat{\theta}^{(1)})) \leq \alpha\bar{\beta}_1 \cdot (\hat{\theta}^{(2)} - \hat{\theta}^{(1)})^2
\end{aligned}$$

where the second inequality holds since the interim allocations are bounded within $[0,1]$, and the last inequality holds by the continuity assumption (Assumption 2).

Moreover, note that the utility $\tilde{U}_{\alpha,z}$ increases at most at a rate η after type $\tilde{\theta}^{(1)}$, while the utility in $\hat{U}_{\alpha,z}$ increases at a rate η within the interval of $(\tilde{\theta}^{(1)}, \hat{\theta}^{(3)})$. Therefore, the gain in utility is at least

$$\begin{aligned}
&(1-\alpha) \cdot \int_{\tilde{\theta}^{(1)}}^{\hat{\theta}^{(3)}} \hat{U}_{\alpha,z}(\theta)\,\mathrm{d}F(\theta) - (1-\alpha) \cdot \int_{\tilde{\theta}^{(1)}}^{\hat{\theta}^{(3)}} \tilde{U}_{\alpha,z}(\theta)\,\mathrm{d}F(\theta) \\
\geq&\,(1-\alpha) \cdot (F(\hat{\theta}^{(3)}) - F(\tilde{\theta}^{(1)})) \cdot (\hat{U}_{\alpha,z}(\tilde{\theta}^{(1)}) - \tilde{U}_{\alpha,z}(\tilde{\theta}^{(1)})) \\
\geq&\,(1-\alpha) \cdot (F(\hat{\theta}^{(3)}) - F(\tilde{\theta}^{(1)})) \cdot (\eta \cdot (\tilde{\theta}^{(1)} - \hat{\theta}^{(1)}) - \epsilon_1) \\
\geq&\,\frac{1}{2\eta}(1-\alpha) \cdot \underline{\beta}_1 \cdot (\eta \cdot (\tilde{\theta}^{(1)} - \hat{\theta}^{(1)}) - \epsilon_1).
\end{aligned}$$

Since the matching efficiency of contest $(\tilde{Q}_{\alpha,z}, \tilde{U}_{\alpha,z})$ is upper bounded by the efficient allocation rule, combining the inequalities, we have that

$$\begin{aligned}
&\mathrm{Obj}_\alpha(\tilde{Q}_{\alpha,z}, \tilde{U}_{\alpha,z}) - \mathrm{Obj}_\alpha(\hat{Q}_{\alpha,z}, \hat{U}_{\alpha,z}) \leq \alpha\bar{\beta}_1 \cdot (\hat{\theta}^{(2)} - \hat{\theta}^{(1)})^2 - \frac{1}{2\eta}(1-\alpha) \cdot \underline{\beta}_1 \cdot (\eta \cdot (\tilde{\theta}^{(1)} - \hat{\theta}^{(1)}) - \epsilon_1) \\
&\leq \alpha\bar{\beta}_1 \cdot \left((\epsilon_2 + \epsilon_0)^2 \cdot \frac{\bar{\beta}_1}{\underline{\beta}_1} + \epsilon_0 + \epsilon_2\right)^2 - \frac{1}{2\eta}(1-\alpha) \cdot \left(\eta\left(\epsilon_2 - \epsilon_0 - \sqrt{\frac{8\epsilon_0\bar{\beta}_1}{\eta\underline{\beta}_1}}\right) - \epsilon_1\right) < 0.
\end{aligned}$$

The last inequality comes from the choice of $\epsilon_0, \epsilon_1, \epsilon_2$. Therefore, the contest $(\tilde{Q}_{\alpha,z}, \tilde{U}_{\alpha,z})$ is not optimal. □

B Non-linear Costs

In the main text we have shown that with linear effort costs, contests are optimal among general mechanisms. In this section, we will show that our result does not rely on the linearity assumption.

Our result extends to convex cost function.

Let $c_i(s_i|\theta_i)$ be the cost of effort for agent i with type θ_i producing a signal s_i.

Assumption 3. *$c_i(s_i|\theta_i) = C_i((s_i - \theta_i)^+) = C_i(e_i)$ for all $e_i \geq 0$.*

Assumption 4. *For any agent i, we have $C_i''(e_i) \geq 0$ and $C_i'''(e_i) \leq 0$.*

Under quadratic effort cost $C_i(e_i) = \frac{1}{2}e_i^2$, Assumption 4 is satisfied.

Lemma 11. *Under Assumption 3 and 4, for any agent i, any $\epsilon \geq 0$, any distribution G supported on $\mathbb{R}_+$ and constant e_G such that $C_i(e_G) = \mathbf{E}_{e\sim G}[C_i(e)]$, we have $C_i(\epsilon + e_G) \geq \mathbf{E}_{e\sim G}[C_i(\epsilon + e)]$, i.e., for any θ, $c_i(\theta + e_G + \epsilon|\theta) \geq \mathbf{E}_{e\sim G}[c_i(\theta + e + \epsilon|\theta)]$.*

Proof. Let $\Delta_i(\epsilon) \triangleq C_i(\epsilon + e_G) - \mathbf{E}_{e\sim G}[C_i(\epsilon + e)]$. Note that by the definition of e_G, we have $\Delta_i(0) = 0$. Therefore, to prove Lemma 11, it is sufficient to show that $\Delta_i'(\epsilon) \geq 0$ for any $\epsilon \geq 0$. Let the expected effort level be $\mu_G \triangleq \int e \,\mathrm{d}G(e)$, since the cost function C_i is convex, we have $e_G \geq \mu_G$. Therefore,

$$\Delta_i'(\epsilon) = C_i'(\epsilon + e_G) - \mathbf{E}_{e\sim G}\big[C_i'(\epsilon + e)\big] \geq C_i'(\epsilon + \mu_G) - \mathbf{E}_{e\sim G}\big[C_i'(\epsilon + e)\big] \geq 0$$

where the inequalities hold since C_i' is increasing and concave for any i by Assumption 4. □

Intuitively, Lemma 11 states that when replacing the stochastic effort recommendation drawn from G by a deterministic effort recommendation, which is its certainty equivalent effort level e_G, each type's cost of effort weakly increases under the deterministic recommendation when they misreport as a higher type. Lemma 11 lies in the heart of the proof for showing the optimality of contests, and Assumption 4 is one sufficient assumption that guarantees this property.

To simplify the exposition in the later analysis, we introduce one more assumption and the following notations.

Assumption 5. *The type space of agent i is discrete and finite, i.e., $\Theta_i = \{\hat{\theta}_i^{(0)}, \ldots, \hat{\theta}_i^{(m)}\}$, with $\hat{\theta}_i^{(0)} < \cdots < \hat{\theta}_i^{(m)}$.*

For any agent i with type θ_i, let his expected utility for generating the signal s_i and receiving allocation x_i be $\tilde{U}_i(\theta_i; x_i, s_i) \triangleq x_i - C_i((s_i - \theta_i)^+)$. For a stochastic mechanism, let D_i be the distribution over the allocation-signal pair, and define the expected utility of agent i with type θ_i as $\tilde{U}_i(\theta_i; D_i) \triangleq \mathbf{E}_{(x_i,s_i)\sim D_i}\Big[\tilde{U}_i(\theta_i; x_i, s_i)\Big]$. Notice that a stochastic mechanism is only implementable by a general mechanism that is not a contest. In contests, it is without loss to focus on deterministic signal recommendation.

Theorem 6. *Under Assumption 3, 4 and 5, for any interim allocation-utility pair $(\boldsymbol{Q}, \boldsymbol{U})$, where $\boldsymbol{Q}$ is monotone, if it is implementable by a general mechanism, then there exists another utility profile $\boldsymbol{U}^\dagger$ such that $(\boldsymbol{Q}, \boldsymbol{U}^\dagger)$ is implementable by a contest. Moreover, $U_i^\dagger(\theta_i) \geq U_i(\theta_i)$ for any agent i, type θ_i.*

Proof. For any agent i and any monotone interim allocation Q_i, we construct the utility function $U_i^\dagger$ by induction. First let $U_i^\dagger(\hat{\theta}_i^{(0)}) = Q_i(\hat{\theta}_i^{(0)})$. For any $k \geq 1$, let $s_i^{(k)}$ be the signal such that $U_i^\dagger(\hat{\theta}_i^{(k-1)}) = \tilde{U}_i(\hat{\theta}_i^{(k-1)}; Q_i(\hat{\theta}_i^{(k)}), s_i^{(k)})$. That is, the agent with type $\hat{\theta}_i^{(k-1)}$ is indifferent between reporting truthfully to receive his expected utility under $U_i^\dagger$ and deviating to the option $(s_i^{(k)}, Q_i(\hat{\theta}_i^{(k)}))$, i.e., generating signal $s_i^{(k)}$ and receiving allocation $Q_i(\hat{\theta}_i^{(k)})$.

By construction, the interim allocation and utility $(\boldsymbol{Q}, \boldsymbol{U}^\dagger)$ coincide with the utility each agent i with type $\hat{\theta}_i^{(k)}$ would receive if he gets allocation $Q_i(\hat{\theta}_i^{(k)})$ as long as he generates a signal $s_i^{(k)}$. It remains to show that this is an equilibrium of a contest, i.e., no agent and no type has strict incentive to deviate (**step 1**); and the corresponding interim utility $U_i^\dagger(\hat{\theta}_i^{(k)})$ is weakly higher than $U_i(\hat{\theta}_i^{(k)})$ (**step 2**).

- **Step 1**: *Incentive compatibility.* By construction, the local incentive compatibility is guaranteed, i.e., each agent and each type has no strict incentive to deviate to an adjacent type. We thus only need to ensure there is global incentive compatibility. This is guaranteed if agent's utility satisfies the standard single-crossing property in the type-allocation space. Indeed, under convex cost function, this property is satisfied for menu options with deterministic signal recommendations.

- **Step 2**: *Higher utilities.* We prove this by induction. First note that $U_i^\dagger(\hat{\theta}_i^{(0)}) = Q_i(\hat{\theta}_i^{(0)}) \geq U_i(\hat{\theta}_i^{(0)})$. For any $k \geq 1$, let $\tilde{s}_i^{(k)}$ be the signal such that $U_i(\hat{\theta}_i^{(k)}) = \tilde{U}_i(\hat{\theta}_i^{(k)}; Q_i(\hat{\theta}_i^{(k)}), \tilde{s}_i^{(k)})$. Let $u_{k-1,k}$ be the expected utility of the agent with type $\hat{\theta}_i^{(k-1)}$ by misreporting as $\hat{\theta}_i^{(k)}$. Notice that $\tilde{s}_i^{(k)} - \hat{\theta}_i^{(k)}$ is the equivalent deterministic effort recommendation, while the effort recommendation for reported type $\hat{\theta}_i^{(k)}$ is potentially stochastic. Thus Lemma 11 implies that $u_{k-1,k} \geq \tilde{U}_i(\hat{\theta}_i^{(k-1)}; Q_i(\hat{\theta}_i^{(k)}), \tilde{s}_i^{(k)})$. Recall that by construction, $\tilde{U}_i(\hat{\theta}_i^{(k-1)}; Q_i(\hat{\theta}_i^{(k)}), s_i^{(k)}) = U_i(\hat{\theta}_i^{(k-1)})$ and by incentive compatibility $U_i(\hat{\theta}_i^{(k-1)}) \geq u_{k-1,k}$. Combining the two, we have that $s_i^{(k)} \leq \tilde{s}_i^{(k)}$. Therefore,

$$U_i^\dagger(\hat{\theta}_i^{(k)}) = \tilde{U}_i(\hat{\theta}_i^{(k)}; Q_i(\hat{\theta}_i^{(k)}), s_i^{(k)}) \geq \tilde{U}_i(\hat{\theta}_i^{(k)}; Q_i(\hat{\theta}_i^{(k)}), \tilde{s}_i^{(k)}) = U_i(\hat{\theta}_i^{(k)}). \qquad \square$$

Theorem 6 implies that contests are optimal among general mechanisms with monotone interim allocations. The monotonicity constraint is a reasonable assumption because in practice, providing higher allocation to lower types may be perceived as unfair.[35] For example, in school admissions,

[35]The monotonicity constraints are imposed similarly in the application of insurance contracts due to practical concerns (Gershkov et al., 2022).

students with higher talent should receive higher probability of being admitted into schools. In government's subsidy programs, people with lower income, or higher financial needs, should receive subsidy from the government with a higher chance.

In Section 8.1, we have shown that there exists non-monotone interim allocation rules that are implementable by general mechanisms. This is because agent's utility does not satisfy single-crossing property in general mechanisms. However, Theorem 1 showed that such non-monotone interim allocation rules are not optimal for linear costs. In Theorem 7, we generalize this idea by showing that under an additional assumption which we call no-concave-crossing, non-monotone interim allocation rules are not optimal.

Assumption 6 (No-Concave-Crossing)**.** *For any agent i, any x_i, s_i and any D_i, if there exist $\hat{\theta}_i < \hat{\theta}'_i$ such that $\tilde{U}_i(\hat{\theta}_i; x_i, s_i) \le \tilde{U}_i(\hat{\theta}_i; D_i)$ and $\tilde{U}_i(\hat{\theta}'_i; x_i, s_i) \le \tilde{U}_i(\hat{\theta}'_i; D_i)$, we have $\tilde{U}_i(\theta_i; x_i, s_i) \le \tilde{U}_i(\theta_i; D_i)$ for any $\theta_i \in [\hat{\theta}_i, \hat{\theta}'_i]$.*

Note that the above condition only requires no-concave-crossing between the utility curve generated by a deterministic recommendation and the utility curve generated by general randomized recommendations. Such no-concave-crossing condition usually is violated if we consider two randomized recommendations. This assumption is satisfied if the cost function is linear or quadratic.

Theorem 7. *Under Assumptions 3, 4, 5 and 6, for any interim allocation-utility pair $(\boldsymbol{Q}, \boldsymbol{U})$ that is implementable by a general mechanism, there exists $(\boldsymbol{Q}^\dagger, \boldsymbol{U}^\dagger)$ with monotone $\boldsymbol{Q}^\dagger$ that is implementable by a contest and attains weakly higher objective value.*

Proof. Similar to the proof of Theorem 1, let $\boldsymbol{Q}^\dagger$ be the monotone rearrangement of $\boldsymbol{Q}$ that is feasible, monotone and weakly improves matching efficiency. We construct $\boldsymbol{U}^\dagger$ as in Theorem 6 such that $(\boldsymbol{Q}^\dagger, \boldsymbol{U}^\dagger)$ is implementable by a contest. It remains to show that $\boldsymbol{U}^\dagger$ weakly improves the agents' utilities for all types under Assumption 6.

Since $\boldsymbol{Q}^\dagger$ is a monotone rearrangement of $\boldsymbol{Q}$, for any θ_i in the support of F_i, there exists another type in the support $\theta'_i \ge \theta_i$ such that $Q_i(\theta'_i) \le Q_i^\dagger(\theta_i)$. Therefore, for any agent i, since U_i is non-decreasing and no greater than the allocation Q_i, we have

$$U_i^\dagger(\hat{\theta}_i^{(0)}) = Q_i^\dagger(\hat{\theta}_i^{(0)}) \ge \min_{\theta_i} Q_i(\theta_i) \ge U_i(\hat{\theta}_i^{(0)}).$$

For any $k \ge 1$, if $Q_i^\dagger(\hat{\theta}_i^{(k)}) \ge Q_i(\hat{\theta}_i^{(k)})$, then similar to Theorem 6, we have $U_i^\dagger(\hat{\theta}_i^{(k)}) \ge U_i(\hat{\theta}_i^{(k)})$ and we are done.

If instead $Q_i^\dagger(\hat{\theta}_i^{(k)}) < Q_i(\hat{\theta}_i^{(k)})$, then there exists $k' > k$ such that $Q_i^\dagger(\hat{\theta}_i^{(k)}) \ge Q_i(\hat{\theta}_i^{(k')})$. Let s_i^k

and $\tilde{s}_i^{(k')}$be the signal such that

$$U_i^\dagger(\hat{\theta}_i^{(k-1)}) = \tilde{U}_i(\hat{\theta}_i^{(k-1)}; Q_i^\dagger(\hat{\theta}_i^{(k)}), s_i^{(k)})$$
$$U_i^\dagger(\hat{\theta}_i^{(k-1)}) = \tilde{U}_i(\hat{\theta}_i^{(k-1)}; Q_i(\hat{\theta}_i^{(k')}), \tilde{s}_i^{(k')}).$$

Note that $Q_i^\dagger(\hat{\theta}_i^{(k)}) \geq Q_i(\hat{\theta}_i^{(k')})$ implies $s_i^{(k)} \geq \tilde{s}_i^{(k')}$.

Let $D_i^{(k)}$ be the distribution over allocation and signal recommendation to type $\hat{\theta}_i^{(k)}$ under the mechanism that gives agent i interim utility U_i. We first show that the following two inequalities hold.

- $\tilde{U}_i(\hat{\theta}_i^{(k')}; Q_i(\hat{\theta}_i^{(k')}), \tilde{s}_i^{(k')}) \geq \tilde{U}_i(\hat{\theta}_i^{(k')}; D_i^{(k)})$. This is because

  $$\tilde{U}_i(\hat{\theta}_i^{(k')}; Q_i(\hat{\theta}_i^{(k')}), \tilde{s}_i^{(k')}) \geq U_i(\hat{\theta}_i^{(k')}) \geq \tilde{U}_i(\hat{\theta}_i^{(k')}; D_i^{(k)})$$

  where the first inequality is implied by Lemma 11 and the second inequality is implied by incentive compatibility.

- $\tilde{U}_i(\hat{\theta}_i^{(k-1)}; Q_i(\hat{\theta}_i^{(k')}), \tilde{s}_i^{(k')}) \geq \tilde{U}_i(\hat{\theta}_i^{(k-1)}; D_i^{(k)})$. This is because

  $$\tilde{U}_i(\hat{\theta}_i^{(k-1)}; Q_i(\hat{\theta}_i^{(k')}), \tilde{s}_i^{(k')}) = U_i^\dagger(\hat{\theta}_i^{(k-1)}) \geq U_i(\hat{\theta}_i^{(k-1)}) \geq \tilde{U}_i(\hat{\theta}_i^{(k-1)}; D_i^{(k)}).$$

  The first inequality holds by the induction assumption. The second inequality holds since $\tilde{U}_i(\hat{\theta}_i^{(k-1)}; D_i^{(k)})$ is type $\hat{\theta}_i^{(k-1)}$'s utility for deviating the report to $\hat{\theta}_i^{(k)}$ and always following the signal recommendation.

Combining two inequalities, since $\hat{\theta}_i^{(k-1)} < \hat{\theta}_i^{(k)} < \hat{\theta}_i^{(k')}$, Assumption 6 immediately implies that $\tilde{U}_i(\hat{\theta}_i^{(k)}; Q_i(\hat{\theta}_i^{(k')}), \tilde{s}_i^{(k')}) \geq \tilde{U}_i(\hat{\theta}_i^{(k)}; D_i^{(k)})$. Moreover, since $Q_i^\dagger(\hat{\theta}_i^{(k)}) \geq Q_i(\hat{\theta}_i^{(k')})$, $s_i^{(k)} \geq \tilde{s}_i^{(k')}$, and the utilities of the agent given these two options coincide at type $\hat{\theta}_i^{(k-1)}$, we have that

$$U_i^\dagger(\hat{\theta}_i^{(k)}) = \tilde{U}_i(\hat{\theta}_i^{(k)}; Q_i^\dagger(\hat{\theta}_i^{(k)}), s_i^{(k)}) \geq \tilde{U}_i(\hat{\theta}_i^{(k)}; Q_i(\hat{\theta}_i^{(k')}), \tilde{s}_i^{(k')}) \geq \tilde{U}_i(\hat{\theta}_i^{(k)}; D_i^{(k)}) = U_i(\hat{\theta}_i^{(k)}). \qquad \square$$

There are two caveats in the results of Theorems 6 and 7. First, our construction only works for distributions with finite support. We have extended the argument for continuous distributions under the setting with linear costs (c.f., Theorem 1). It is not clear whether such a general argument exists for convex costs. Second, our assumption of no-concave-crossing (Assumption 6) is only sufficient but not necessary for showing that contests are optimal among non-monotone mechanisms. Moreover, this assumption is hard to interpret in practice. We conjecture that there exist weaker and interpretable conditions for ensuring the optimality of contests. We leave these as interesting open questions for future follow-up works.